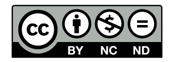
Análisis de contenidos

Manela Juncà Campdepadrós

PID_00195714





Los textos e imágenes publicados en esta obra están sujetos –excepto que se indique lo contrario– a una licencia de Reconocimiento-NoComercial-SinObraDerivada (BY-NC-ND) v.3.0 España de Creative Commons. Podéis copiarlos, distribuirlos y transmitirlos públicamente siempre que citéis el autor y la fuente (FUOC. Fundación para la Universitat Oberta de Catalunya), no hagáis de ellos un uso comercial y ni obra derivada. La licencia completa se puede consultar en http://creativecommons.org/licenses/by-nc-nd/3.0/es/legalcode.es

Índice

Introducción

El objetivo del análisis de contenido es identificar y representar de manera precisa la materia de los documentos, con el objetivo de permitir la recuperación. Esta parte del análisis documental establece los puntos de acceso por materias o contenidos de los documentos.

Se basa en dos operaciones:

a) El **resumen**, que es la representación abreviada y precisa del contenido.

b) La **indización**, que consiste en representar el contenido del documento mediante términos de indización extraídos de **lenguajes documentales**: notaciones, encabezamientos de materias, descriptores, identificadores, palabras clave, unitérminos. Cuando se representa el contenido siguiendo un sistema de clasificación en lugar de una *indización* se conoce como *clasificación*.

Las **normativas** que usamos en esta parte del análisis documental son:

- UNO 50-103-90, preparación de resúmenes.
- UNO 50-121-91, métodos para el análisis de documentos, determinación de su contenido y selección de términos de indización.
- Las normativas propias de cada lenguaje documental: vocabulario, combinaciones, mantenimiento, actualización.

Campos propios del análisis de contenido en la referencia de Pierre Bonnassie: materia y resumen

Campos propios del análisis de contenido en la referencia de Pierre Bonnassie

Materia	Historia medieval - Terminología
Resumen	Este es un libro poco corriente. Ni diccionario ni manual, significa una nueva y eficaz forma de introducción –a la vez analítica y sintética– a los problemas de la historia de la Edad Media. En efecto, a partir del análisis de medio centenar de conceptos fundamentales y de su evolución semántica, el profesor Pierre Bonnassie, de la Universidad de Toulouse, consigue definir, con insólita precisión, las grandes cuestiones que hoy tiene planteadas la historia medieval. El resultado es un texto innovador, de uso obligado para profesores y estudiantes, que encontrarán en él un instrumento de trabajo insustituible.

Los lenguajes documentales usados tradicionalmente en los archivos son cuadros de clasificación construidos a medida del fondo. El análisis de contenido es sintético, no se analizan los documentos individualmente, sino el fondo en su conjunto o los expedientes, dado que un documento forma parte de una cadena de documentos ordenados (cronológicamente, orgánicamente, funcionalmente) y aislado pierde su contexto. La clasificación puede ser orgánica, funcional (por funciones, por grandes materias) o mixta. No obstante, para describir el contenido de un expediente o de una serie, más allá de sus

Clasificación orgánica

La clasificación orgánica es el retrato de la estructura orgánica de la entidad que haya generado la documentación.

funciones o situación orgánica, hay lenguajes documentales, como los tesauros, que permiten identificar las temáticas para la posterior explotación de la información contenida en los documentos.

En bibliotecas y centros de documentación se usan la mayoría de los lenguajes documentales. Los más habituales son los sistemas de clasificación, como la Clasificación Decimal Universal (CDU) o la Clasificación Dewey, los listados de autoridades, las listas de encabezamientos de materia, los tesauros y la indización automática por palabras clave.

CDU

CDU es la sigla de *Clasificación Decimal Universal*.

En este módulo veremos con detenimiento las técnicas de resumen y los lenguajes documentales, como instrumentos para describir el contenido de los documentos.

1. El resumen humano y automático

Según la norma UNE 50-103-90 *Preparación de resúmenes*, un **resumen** es la presentación abreviada y precisa de un documento, sin interpretación ni crítica y sin mención expresa del autor del resumen.

Ved también

Encontraréis la norma UNO 50-103-90 en el espacio "Materiales y fuentes" de las aulas.

Cuando decimos documento nos estamos refiriendo a todo tipo de documento, sea cual sea su soporte material. Podemos resumir un texto, la imagen de una fotografía, un vídeo, audios, información en línea o hipertextos, un expediente o una serie.

Los resúmenes, como la indización, pueden ser de elaboración humana o automática. En el primer caso hay cuatro tipos de personas que pueden redactar un resumen. En el caso de los resúmenes automáticos, se trata de un software.

1) **Resumen humano**:

a) El **autor** del documento. Los resúmenes elaborados por los propios autores son muy habituales en el mundo de las comunicaciones científicas y tecnológicas.

b) Un **especialista** en la materia de la que trata el documento.

c) La **editorial**. Son los resúmenes que aparecen en la contraportada de los libros impresos y que tienen una función claramente publicitaria.

d) Un **profesional de la documentación**. Aporta su conocimiento sobre la redacción de buenos resúmenes y los elabora pensando en las utilidades futuras.

2) **Resumen automático**: los programas se conocen como programas resumidores de textos o *Automatic Text Summarizer*.

Programas resumidores de textos

Un ejemplo de programas resumidores de textos es Swesum, que hace un análisis estadístico del texto y elabora el resumen con los fragmentos que contienen las palabras más ponderadas (más repetidas pero con significado).

La norma internacional ISO 214:1976, traducida por AENOR como norma UNE 50-103-90 *Preparación de resúmenes*, establece las directrices que se tienen que seguir para presentar los resúmenes en los documentos. Pone especial énfasis en la preparación de resúmenes por parte de los autores de los documentos primarios y en la misma publicación.

Redactar un resumen es fácil. Lo difícil es redactar un buen resumen. El punto de inflexión es la calidad del resumen, que lo hará más o menos útil en un sistema documental. Un resumen propagandístico no aportará muchos conceptos principales para indizar, aunque haya sido un buen reclamo para las ventas.

Ejemplo de resumen elaborado por la editorial con finalidad publicitaria

Sagan, Carl. *Cosmos*. Traducció: Albert Santamaria i Martínez; pròleg: Ricard Guerrero. Barcelona: Publicacions i Edicions de la Universitat de Barcelona: Omnis Cellula, cop. 2006.

"He aquí una de las obras más destacadas de la literatura internacional de divulgación científica, publicada por primera vez en catalán. Una obra imprescindible de uno de los grandes maestros de la divulgación, que nos introduce en los grandes enigmas que la humanidad ha tratado de entender y explicar desde tiempos inmemoriales, y por los cuales ha nacido lo que llamamos ciencia.

Desde la infinitud del Universo hasta el mundo invisible de los átomos, desde el nacimiento de las estrellas hasta la aparición de la vida, Carl Sagan consigue transmitir los conocimientos de la ciencia actual de una manera clara y apasionante."

Para un analista sólo tendría utilidad el último párrafo, en qué aparecen términos como *universo*, *átomos*, *estrellas*, *vida*.

El resumen es útil en la fase de descripción y es un excelente instrumento de recuperación, ya que el resumen ofrece más datos que la simple referencia documental. La principal utilidad del resumen es la de difundir la información.

Difundir la información

Cada vez más bases de datos referenciales ofrecen el resumen de sus monografías y revistas, como por ejemplo Ebsco, Dialnet, Compludoc, CBUC, Eric database o *ISI current contents connect*. También lo hacen las bases de datos de novedades editoriales, por ejemplo la editorial Trea (recomendamos el acceso desde la biblioteca de la UOC).

Además, el resumen tiene otras utilidades, tal como dice la norma UNE 50-103-90:

a) Determinar la pertenencia: un resumen bien elaborado capacita a los lectores para identificar de forma rápida y precisa el contenido de un documento y decidir si hay que leerlo en su totalidad.

b) Evitar la lectura del texto completo en documentos de interés secundario. Un resumen bien elaborado proporciona suficiente información sobre temas que no sean de interés principal para el lector. Ahorra tiempo al usuario.

c) Ayudar en la búsqueda automatizada. Los resúmenes automatizados incorporados en los catálogos son muy útiles para:

- Extraer términos de indización de su texto, es decir, indizar a partir del resumen.

- Hacer búsquedas de palabras clave que no se encuentran en el título.

- Servir de control bibliométrico, al comparar los términos usados en una ecuación de búsqueda con los términos que aparecen en un resumen y así establecer la pertinencia de la recuperación.

- Ayudar a la difusión desde los servicios de alerta.

Según María Pinto (1992), las **características de un resumen** son las siguientes:

- Brevedad. Se tienen que omitir datos preliminares o temas del conocimiento común.

- Pertinencia. El resumen se tiene que adecuar al mensaje principal del documento, sin obviar o interpretar los datos.

- Claridad y coherencia. Frases completas, dotadas de coherencia lineal y global.

- Profundidad. Varía en función del tipo de resumen o de los diferentes niveles de detalle que se persigan.

- Consistencia lingüística. Un resumen se tiene que adaptar a las pautas lingüísticas en uso y tiene que tener en cuenta las reglas morfológicas y sintácticas correspondientes.

- Proximidad cronológica entre las ediciones del documento original y el resumen. Es importante que el tiempo transcurrido entre la publicación del original y el resumen no sea excesivo, especialmente en ámbitos científicos y técnicos.

Lectura complementaria

Podéis ampliar la información sobre el resumen leyendo la obra siguiente:

M. Pinto Batanea (1992). *El resumen documental: principios y métodos*. Madrid: Pirámide/Fundación Germán Sánchez Ruipérez (Biblioteca del Libro, Y).

A modo de conclusión

- El resumen es la presentación abreviada y precisa de un documento, sin interpretación ni crítica y sin mención expresa del autor del resumen.

- El resumen puede ser redactado por el autor del documento, un especialista en la materia, la editorial, un documentalista o un programa informático.

- El resumen es útil en dos fases de la cadena: en los procesos de selección y adquisición que se da en la primera fase de la cadena y en la fase de salida, donde es un excelente instrumento de recuperación.

- La principal utilidad del resumen es la de difundir la información, pero además, el resumen tiene otras utilidades, como determinar la pertinencia, evitar la lectura del texto completo en documentos marginales y ayudar a la búsqueda automatizada.

- Los resúmenes automatizados incorporados en los catálogos son muy útiles para extraer términos de indización del texto, para hacer búsquedas de palabras clave que no se encuentran en el título, para servir de control bibliométrico y ayudar a la difusión a través de los servicios de alerta.

1.1. Tipos de resúmenes

Hay diversos tipos de resúmenes, según el tamaño, los usuarios y la profundización en el contenido. Los tipos más habituales son los resúmenes informativos, indicativos y selectivos.

1) Resumen informativo

Redactaremos el tema central, temas adicionales, naturaleza y objetivo del documento, metodología, resultados, conclusiones y anexos. La idea de fondo es que un resumen informativo puede sustituir en ocasiones la lectura del documento original. La norma UNE 50-103-90 recomienda que el esquema a seguir sea el de:

objetivo + metodología + resultados (o conclusiones)

Sin embargo, no hay que seguir forzosamente este orden, ya que hay entornos, como el técnico científico, donde se prefieren los resúmenes orientados a los resultados (para que la discriminación sea más rápida).

En cuanto al tamaño del resumen, la norma da pautas pero advirtiendo que el contenido del documento es más significativo que las pautas para determinar la extensión del resumen. De todas maneras la norma nos sugiere:

- Monografías, informes, tesis: 500 palabras.
- Artículos de revista, capítulos de monografías: 250 palabras.
- Comunicaciones breves: 100 palabras.

Ejemplo de resumen informativo

Consuegra Fernández, Jesús: "El Ajedrez: evolución y claves de un juego milenario". En *Mundo antiguo*. Madrid: 2002. n° 3-4, año 1, p. 60-61.

"Artículo divulgativo sobre el juego del ajedrez, estructurado según sus orígenes, antigüedad, expansión, variantes y simbolismo.

El origen del ajedrez es hindú y el primer representante conocido es el Ghaturanga, aparecido entre el 3000 y el 2000 a.C. en Sri Lanka, aunque no aparece documentado hasta el siglo VII d.C.

Del Ghaturanga proceden en cascada las diferentes variantes del ajedrez: de la India viajó a Persia en el siglo VI d.C., donde pasó de los 4 jugadores originales a 2 en la versión persa Shatranj. Desde Persia se extendió hacia Occidente y hacia Oriente.

Hacia Occidente: paralela a la expansión árabe, el juego llega a la Península Ibérica durante la Alta Edad Media, y desde aquí se expande al resto de Europa y al resto del mundo en la época de las colonizaciones.

Hacia Oriente: en la China, en el s. VII d.C., el ajedrez toma la forma del ajedrez chino Xiang qi; en el Japón, el Shogi; en Indochina, el ajedrez birmano y tailandés. Tanto en Oriente como en Occidente, el ajedrez presenta innumerables variaciones locales.

El tablero y las fichas parecen poseer un significado simbólico. El tablero, con la alternancia de casillas blancas y negras, forma un mandala. El simbolismo de las fichas es menos esotérico y ha ido cambiando según los tiempos: obispos, elefantes, etc.

El autor concluye que el ajedrez, además de un juego, es una herramienta educativa de primer orden, casi una ciencia."

Como podéis comprobar, este resumen tiene 237 palabras.

2) Resumen indicativo

Redactaremos sólo las ideas centrales del documento. Su lectura no puede sustituir la lectura del original. Como su nombre sugiere, el resumen indicativo presenta de forma abreviada y muy sintética el contenido o la tipología del documento. Su extensión puede oscilar entre una frase o 4 líneas de texto.

Ejemplo de resumen indicativo

Consuegra Fernández, Jesús: "El Ajedrez: evolución y claves de un juego milenario". En *Mundo antiguo*. Madrid: 2002. n° 3-4, año 1, p. 60-61.

"Artículo divulgativo sobre el juego del ajedrez, trata de su origen hindú, antigüedad, expansión histórica tanto en Oriente como en Occidente, variantes nacionales y simbolismo del tablero y las fichas."

3) Resumen selectivo

Redactaremos sólo una parte concreta del documento. El más habitual es el resumen de conclusiones, pero también hay otros tipos, como la reseña (*review*), que es un análisis del documento con elementos críticos. Este tipo de resumen se adapta muy bien a las necesidades de los usuarios, por ejemplo investigadores o técnicos que necesitan un dato muy concreto sobre el objetivo del documento o las conclusiones a las que llega.

Ejemplo de resumen selectivo

Consuegra Fernández, Jesús: "El Ajedrez: evolución y claves de un juego milenario". En *Mundo antiguo*. Madrid: 2002. n° 3-4, año 1, p. 60-61.

"El ajedrez, además de un juego, es una herramienta educativa de primer orden, casi una ciencia."

A modo de conclusión

Los resúmenes más habituales son el resumen informativo, el indicativo y el selectivo:

- El **resumen informativo** consigna el tema central, temas adicionales, naturaleza y objetivo del documento, metodología, resultados, conclusiones y anexos. La idea de fondo es que un resumen informativo puede sustituir en ocasiones a la lectura del documento original.
- El **resumen indicativo** consigna sólo las ideas centrales del documento. Su lectura no puede sustituir a la lectura del original.
- El **resumen selectivo** consigna sólo una parte concreta del documento. El más habitual es el resumen de conclusiones, pero también hay otros tipos, como la reseña (*review*).

1.2. Resumen automático

Una de las necesidades más perentorias ante el aumento de información digital debido al crecimiento exponencial de Internet es manejar y filtrar el gran volumen de información. Una de las soluciones aportadas por el PLN han sido los programas de resumen automático, que actúan sobre textos, imágenes, webs y correo electrónico.



Hans Peter Luhn

Los primeros en trabajar en el campo de la automatización de los resúmenes fueron Hans Peter Luhn en el año 1958 y Edmundson en 1969, que aplicaron técnicas como la frecuencia de las palabras, o la posición de una frase dentro de un documento para redactar resúmenes sin intervención humana.

A partir de estas primeras investigaciones se han perfeccionado muchas técnicas diferentes basadas en conocimiento y recursos lingüísticos (como las de Lin y Hovy, 2002; Gotti *et al.*, 2007) o basadas en métodos estadísticos y de aprendizaje automático (Hirao *et al.*, 2002; Svore, 2007) (autores citados en Lloret *et al.*, 2008; y Mateo *et al.*, 2003).

Últimamente las investigaciones giran en torno al resumen multidocumento, es decir, resumir más de un documento (Goldstein *et al.*, 2000; Qiu, 2007; Huo y Chen, 2008) de contenidos afines o redundantes (autores citados en Lloret *et al.*, 2008; y Mateo *et al.*, 2003).

Lecturas complementarias

Podéis consultar los resultados de las investigaciones de estos autores en los artículos siguientes:

E. Lloret; O. Ferrández; R. Muñoz; M. Palomar (2008). "Integración del reconocimiento de la impliación textual en tareas automáticas de resúmenes de textos". *Procesamiento del lenguaje natural*, n°. 41, pág. 183-190.

P. L. Mateo; J. C. González; J. Villena; J. L. Martínez (2003). Un sistema para resumen automático de textos en castellano.

Los resúmenes automáticos se conocen también como *extracts*. La terminología anglosajona diferencia así los *extracts* y los *abstracts*. Los *extracts* son los resúmenes formados a partir de la extracción de algunas frases del texto previamente seleccionadas por un programa, mientras que los *abstracts* son los resúmenes elaborados por una persona.

La base de todas las técnicas de funcionamiento de un programa de resúmenes automático es el cómputo de la frecuencia de las palabras.

Hay diversas herramientas para hacer estos cálculos, por ejemplo WVTool. Se trata de contar cuántas veces sale una palabra no vacía en el texto.

Ejemplo de funcionamiento de un programa de resúmenes automático (extraído de Lloret *et al.*, 2008)

"Tropical storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night. There were no reports of casualties."

Oración 1:	Tropical (2) storm (6) Gilbert (7) formed (1) in (0) the (0) eastern (1) Caribbean (1) and (0) strengthened (1) into (0) a (0) hurricane (7) Saturday (4) night (2).
Oración 2:	There (0) were (0) no (0) reports (1) of (0) casualties (1).

Lo primero que vemos es que las palabras vacías, es decir, las palabras que no tienen significado (preposiciones, artículos, verbos) no se computan.

Al lado de cada palabra con significado vemos el número de veces que sale en todo el texto. Se suman los valores, de manera que la oración 1 tiene 3,2 puntos y la oración 2, 0,2. El programa seleccionará la frase 1 como más representativa para el resumen automático.

Este sistema de resumir a partir de las frases con las palabras más significativas en el texto parece simplista pero tiene cierta justificación. Según Kupiec *et al.* (1995) aproximadamente el 80% de las frases en resúmenes humanos están copiadas literalmente o con pequeñas modificaciones del texto original.

A partir de esta base estadística se incorporan otras técnicas para dotar al programa de más conocimiento y paliar la escasa coherencia del resultado, como puede ser, por ejemplo, la resolución de la anáfora o aplicar programas (por ejemplo, WordNet) que proporcionen relaciones como las de sinonimia o hiperonimia, o mecanismos para detectar y eliminar la redundancia.

Definimos brevemente qué son las anáforas y la hiperonimia:

a) Las **anáforas** son la relación de referencia entre un elemento lingüístico y otro anterior en el discurso.

b) Decimos que una palabra es **hiperónima** cuando tiene un campo significativo que incluye otro de menor extensión.

Los expertos consideran que la tecnología actual no tiene problemas para detectar las frases con más significado, pero sí para ordenarlas según su importancia.

Los programas funcionan a grandes rasgos de la siguiente manera: se copia el texto a resumir o bien se escribe la dirección del documento. Se escoge el tipo de documento (académico, periodístico, etc.) y el tanto por ciento de reducción del texto.

A continuación tenéis unos cuantos programas de los más conocidos:

- Connexor
- Daedalus

Anáfora

"El Salón del Hobby ha tenido más de 60.000 visitantes este año. Este salón se ha convertido en la feria de ocio familiar más visitada".

En este ejemplo, la anáfora se da en "este salón", que hace referencia al Salón del Hobby, expresado en la frase anterior. Como se puede comprobar, si en el resumen automático aparece sólo la segunda frase, el lector no sabrá a qué salón hace referencia.

Hiperonimia

Color es un hiperónimo. Su contrario es hipónimo: *amarillo*, *naranja*, *verde* son hipónimos.

- Extractor
- FociSum
- InTEXT (Dynamic Summarizing)
- Inxight Summarizer
- IslandInText
- K-Site de Daedalus
- Pertinence Summarizer
- Sinope Summarizer
- Summarizer
- SweSum[1]
- System Q
- TextAnalyst
- Trestle

(1)Podéis practicar con el programa Swesum, que es gratuito y traduce al español.

El programa K-Site de Daedalus

De entre los programas de resumen automático mencionados, veamos el funcionamiento del programa K-Site de Daedalus. Este programa tiene cinco módulos:

- **Módulo 1: Análisis morfosintáctico**. En este módulo se determina la categoría léxica de cada palabra: sustantivo, verbo, adjetivo, artículo, preposición, etc. También se determina el lema. Estas operaciones permiten distinguir las palabras con significado (sustantivos, adjetivos, verbos) de las vacías (artículos, preposiciones, pronombres, etc.). El lema permite agrupar todas las palabras que son flexiones de otra (info/informar/información/informador/informacional/etc.). El producto final es un listado con las palabras puntuadas y un listado de frases candidatas.

- **Módulo 2: Ponderación de frases**. Este módulo recibe las palabras etiquetadas por el módulo anterior, y su función es escoger entre todas las frases candidatas. Para hacerlo se ayuda de diversos submódulos que ponderan las frases según los parámetros siguientes: la frecuencia, la presencia de palabras indicativas (buscan palabras como *importante, esencial, conclusiones*, etc.), buscan frases que contengan palabras que aparezcan en el título, o que tengan nombres propios, o que la tipografía sea destacada (negritas, cursivas, tamaño superior, etc.) y seleccionan frases que aparezcan en posiciones destacadas en el texto (al principio de cada párrafo, al final a modo de conclusión).

- **Módulo 3: Detección de anáforas**. Una vez tiene las frases seleccionadas, puede ser que se dé el caso de anáforas mal resueltas (una frase contiene una anáfora que se encontraba en la frase previa y que no ha sido seleccionada). El programa busca las anáforas (especialmente los demostrativos pronominales o pronombres personales, por ejemplo *este, aquel, lo que*, *eso*) y su posición en la frase: al principio, entre las seis primeras palabras, en otras posiciones.

- **Módulo 4: Selección de frases**. Este módulo computa toda la información recogida en las fases anteriores: frases candidatas, puntuaciones, detección de anáforas. Selecciona las frases candidatas de puntuación más alta hasta llegar al tanto por ciento pedido por el usuario. Si entre estas frases hay alguna que contenga una anáfora, se selecciona la frase anterior (que contiene la palabra a la cual se está haciendo referencia) siempre y cuando forme parte de las frases candidatas y no sobrepase la longitud del resumen.

- **Módulo 5: Postprocesado del extracto**. Su función es detectar expresiones que conectan partes del texto, ya sea para mostrar causalidad, contraposición, etc. Son expresiones del tipo *por lo tanto, en contra*, etc. Como en el caso de las anáforas, si forman parte de una frase seleccionada, se procura incluir en el resumen la frase con la cual están relacionadas.

Por último, debemos recordar que algunos procesadores de textos, como Microsoft Word, también ofrecen esta opción (*Autosummarize* o Auto-resumen).

A modo de conclusión

- Los resúmenes automáticos (*extracts*) son una de las soluciones aportadas por el PLN para hacer frente al manejo de grandes volúmenes de información en línea.
- Los primeros en trabajar en el campo de la automatización de los resúmenes fueron Hans Peter Luhn en el año 1958 y Edmundson en 1969.
- Las técnicas han evolucionado de los primeros cómputos sobre la frecuencia de las palabras, o la posición de una frase dentro de un documento, a las técnicas basadas en conocimiento y recursos lingüísticos o en métodos estadísticos y de aprendizaje automático.
- La base de todas las técnicas es el cómputo de la frecuencia de las palabras. A partir de esta base estadística, se incorporan otras técnicas para dotar al programa de más conocimiento y paliar la escasa coherencia del resultado, por ejemplo la resolución de la anáfora o se aplican programas que proporcionen relaciones como las de sinonimia o hiperonimia o mecanismos para detectar y eliminar la redundancia.
- Los expertos consideran que la tecnología actual no tiene problemas para detectar las frases con más significado, pero sí para ordenarlas según su importancia.

2. La indización y la recuperación: lenguajes documentales y lenguaje natural

> "Indizar es la acción de describir o identificar un documento con relación a su contenido."
>
> Norma UNE 50-121-91.

Indizar es el resultado de examinar el documento, seleccionar los conceptos y almacenarlos en una base de datos.

Esta definición implica tres acciones, de las cuales la más significativa es la selección de los conceptos y su traducción al lenguaje documental.

Al igual que se ha tratado en el resumen, la indización la puede realizar una persona o un programa.

Si la indización es intelectual, es decir, la llevan a cabo personas, estas personas pueden ser:

- **Profesionales** (documentalistas), que llevan a cabo la tarea de indización de manera individual o en equipo. A su vez, los equipos pueden indizar de manera centralizada o coordinada.

- **Amateurs** (usuarios de Internet que indizan de manera social o *tagging*, por ejemplo, en Delicious).

El elemento humano permite un análisis más rico del documento, captando conceptos y matices que un programa no llegaría a detectar, pero tiene el inconveniente del tiempo que se tiene que dedicar y la coherencia entre indizadores.

La indización automática se realiza a través de un programa informático. Su funcionamiento es muy sencillo: extrae del título, resumen o texto completo las palabras más significativas. Es un método económico y muy rápido.

La recuperación

La recuperación es un proceso paralelo a la indización.

Si se busca un dato concreto, como un título (Hamlet, web semántica) o un autor (Shakespeare, Lluís Codina), la búsqueda no reviste ninguna dificultad, ya que la petición se efectúa con unos datos objetivos y la respuesta solo puede ser "tengo resultados o no tengo resultados". En cambio, cuando no se busca

por un dato concreto sino por un tema, entonces entran en juego las mismas tres fases (examen, selección y traducción) que en la indización, pero con la diferencia de que lo que se examina y se selecciona es la petición del usuario.

1) Examinar la petición del usuario para identificar el contenido.

2) Seleccionar los conceptos principales de la petición.

3) Traducir a un lenguaje documental.

En la recuperación, una de las claves es conocer bien el lenguaje documental que debemos consultar, porque si es así podremos llevar a cabo búsquedas más precisas, sobre todo en el caso de lenguajes controlados (por las relaciones semánticas que establecen entre los términos). Así pues, el primer paso será averiguar qué tipo de indización se encuentra tras la caja de búsqueda.

Los lenguajes documentales que hay tras una fuente de información no son evidentes, tienden a la invisibilidad. Los programas prefieren pantallas de búsqueda muy simples (por ejemplo, Scirus), donde aparece una caja en blanco: sencillo y amigable para el usuario, pero a nosotros no nos puede pasar por alto que esconde un lenguaje documental o, más probablemente, una combinación de lenguajes.

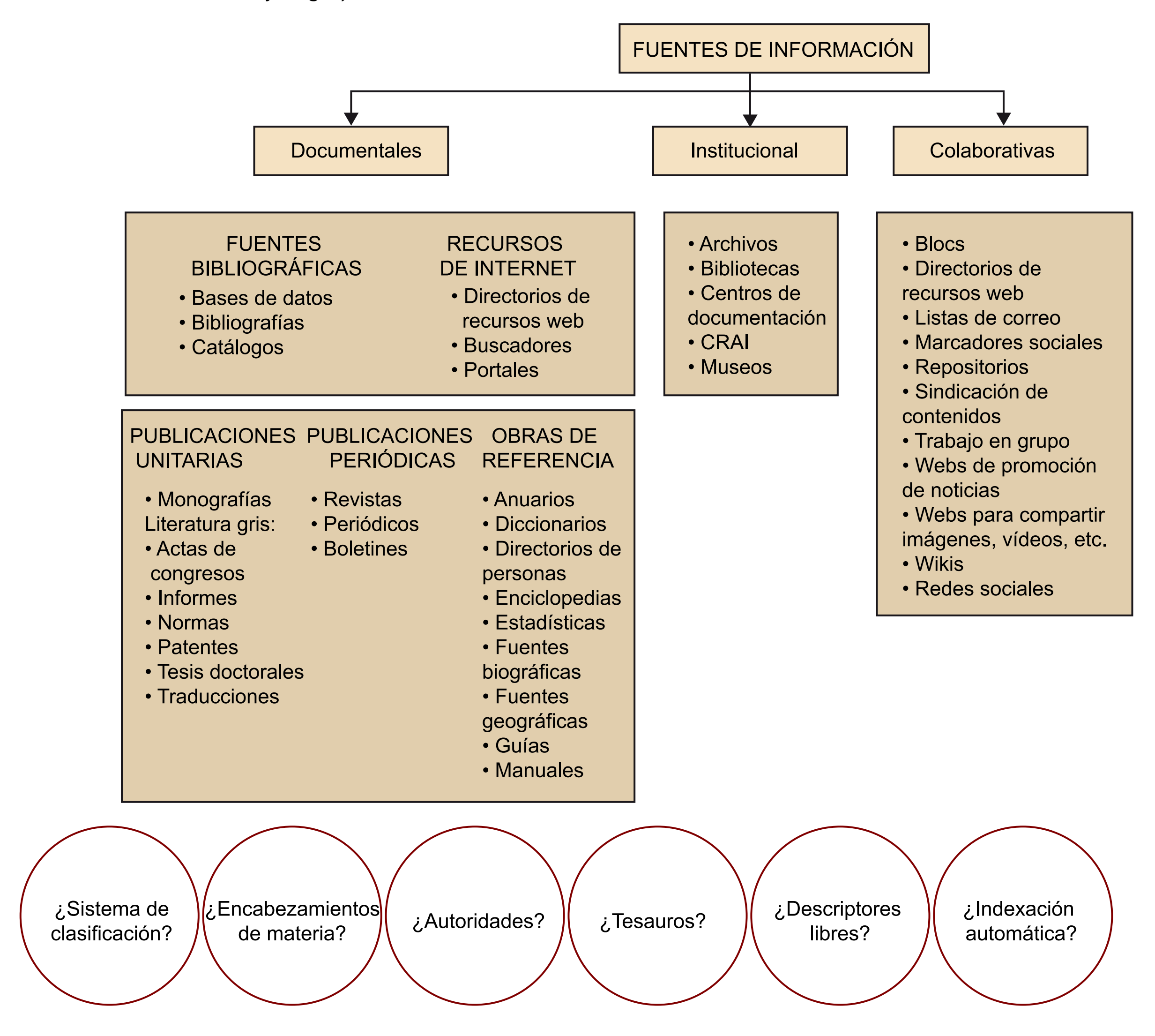
En el proceso de búsqueda probablemente pasaremos de una fuente de información a otra y, en consecuencia, de un tipo de indización a otro.

Mientras la búsqueda se lleve a cabo en buscadores, la indización será **automática y libre**, pero cuando entramos en intranets y bases de datos, la indización cambiará, probablemente, a una **controlada**, en cuyo caso deberemos saber qué tipo de lenguaje las controla.

Ejemplo

Usamos un buscador general como Google (indización automática) para llegar a la web de la Biblioteca de Catalunya y a su catálogo, que está clasificado con CDU, LEMAC y LENOTI (tres lenguajes controlados).

Figura 1. Fuentes de información y lenguajes documentales.



Observación

No se puede diseñar una tabla que relacione el tipo de fuentes de información y el lenguaje que utilizan porque, a pesar de que se sigue cierta tendencia, no son siempre iguales.

Las fuentes de información más estándares son los **catálogos bibliotecarios** (que suelen estar indizados con sistemas de clasificación, listas de encabezamientos de materia y listas de autoridades) o de archivos, y los **buscadores**, que no podrían existir sin la indización automática. Ahora bien, el resto es muy diverso, de modo que podemos llegar a encontrar bases de datos indizadas por tesauros (Unesco) o, simplemente, por descriptores libres (Delicious).

Para saber qué lenguaje indiza la fuente, es útil observar si lleva un menú de opciones con enlaces del tipo "Normalización", "para profesionales", o incluso directamente LEMAC[2] o LCSH[3], es decir, el nombre del lenguaje, irreconocible para un profano pero perfectamente reconocible para los documentalistas.

(2)Lista de encabezamientos de materia en catalán

(3)Library of Congress Subject Headings

En segundo término, podemos reconocer el lenguaje:

- por la forma del término (un código será una clasificación, dos palabras separadas por guión será un encabezamiento de materia);
- por un número de términos en plural (nos dice que se trata de descriptores, habrá que averiguar si son controlados –de un tesauro– o libres –descriptores libres o *tags*);
- por el tipo de fuente (un catálogo o un buscador usan siempre el mismo tipo de lenguaje);
- por la institución que hay tras él;
- por la experiencia del documentalista.

2.1. Lenguaje natural y lenguaje documental

Para indizar necesitamos los lenguajes documentales. ¿Qué diferencia hay entre el lenguaje natural y el documental?

Por **lenguaje natural** entendemos el lenguaje que usamos de forma cotidiana: catalán, castellano, vasco, gallego, francés, etc.

Por **lenguaje documental** entendemos el listado o vocabulario de términos que usamos para indizar y que puede estar en formato libre o controlado.

¿Y por qué hay que controlar los términos del lenguaje natural? Porque el lenguaje natural es ambiguo, los conceptos se pueden representar de formas diversas, dando lugar a problemas de recuperación. El lenguaje natural es rico en terminología, en formas (plurales y singulares), tiempos verbales, acrónimos, sinónimos, polisemias, etc.

La riqueza del lenguaje natural

- Ejemplos de sinónimos del mismo concepto: Cosmos / Universo / Infinito / Firmamento / Cielo.
- Ejemplo del mismo concepto en formas diferentes, siglas o frases, y en idiomas diferentes: OTAN / NATO / Organització del Tractat de l'Atlàntic Nord / Organización del Tratado del Atlántico Norte / North Atlantic Treaty Organization.
- Ejemplo de polisemia: Banco / Planta / Carta / Sierra / Estrella / Lengua / Capital.

La principal diferencia entre el lenguaje natural y el documental controlado es precisamente el control terminológico, que permite representar los conceptos de forma unívoca, sin ambigüedades.

Univocidad

La univocidad consiste en representar un concepto con un único término.

Para ser más concretos, las diferencias se dan en el número de términos del vocabulario, el control de las formas, el control del significado y las relaciones de significado entre términos.

2.1.1. Número de términos

Los lenguajes documentales son entrópicos (Blanca Gil, 2004, pág. 20), es decir, tienden a la selección, a la restricción del vocabulario. Es el proceso contrario del lenguaje natural, que tiende a la abundancia, a la reiteración de conceptos, a la sinonimia en beneficio de una expresión más rica.

Los lenguajes documentales reducen considerablemente el número de términos del lenguaje natural, ya que sólo tienen en consideración los sustantivos y algunos sintagmas nominales, pero no adjetivos, preposiciones, conjunciones, adverbios, verbos, etc. Además, entre todos los sustantivos, escogen uno que representará al resto cuando el significado sea el mismo. Y entre diversas formas aceptadas por el mismo término, sólo una será la aceptada, como es el caso de las siglas.

Los lenguajes documentales son en esencia sencillos, su eficacia aumenta a medida que las reiteraciones y la redundancia son controladas en una única forma que reúne conceptos afines.

2.1.2. Control de las formas

Los lenguajes documentales controlan las formas plural/singular, el uso de acrónimos y siglas y la construcción de las frases, y de esta manera establecen unos modelos.

Modelo	Ejemplo
Sustantivo	Pintura
Sustantivo + adjetivo	Pintura medieval
Sustantivo + preposición + sustantivo	Pintores de vitrales

Estas reglas gramaticales y sintácticas unifican las palabras seleccionadas y las frases.

Ejemplos en las listas de encabezamientos de materia

- Se acostumbra a usar el singular para expresar conceptos abstractos. Así, por ejemplo, es *solidaridad* y no *solidaridades*.
- No se permite el uso de siglas; se prefiere la expresión entera del concepto y en la lengua del servicio de información y documentación (SID[4]). Por ejemplo, Organización del Tratado del Atlántico Norte.
- Es preferible la expresión natural del concepto compuesto, y no su forma inversa. Es correcto *Objetos de arte*, y no *Arte, objetos de*.

(4)SID es la sigla *de servicio de información y documentación.*

2.1.3. Control del significado

Los problemas más importantes en cuanto al significado son la sinonimia y la polisemia.

a) Sinonimia: decimos que las palabras son sinónimas cuando tienen el mismo significado. En un sistema documental, si no se controlan y se usan indiscriminadamente, comportan silencio documental. En el caso de "alimento, nutriente, comida, provisión", el usuario puede estar buscando por "alimento" y no recuperar documentos porque se encuentran indizados con otras formas, como "nutriente". La solución de los lenguajes controlados es recoger todos los términos sinónimos y seleccionar uno para representar a todo el conjunto de términos que tienen el mismo significado, porque dos sinónimos son sustituibles el uno por el otro en cualquier contexto.

Ejemplo

Una lista de encabezamientos de materia como la del Consejo Superior de Investigaciones Científicas (CSIC) recoge todos estos sinónimos:

- Hispanoamericanos.
- Iberoamericanos.
- Latinoamericanos.
- Sudamericanos.

Pero sólo da como término aceptado "Latinoamericanos". Si al SID[5] llegara un documento titulado "Los sudamericanos del siglo XX", el analista lo indizaría como **Latinoamericanos**, ya que es el término aceptado.

(5) A partir de ahora denotamos *servicios de información y documentación* con la sigla SID.

b) Polisemia: decimos que dos palabras son polisémicas cuando el mismo signo lingüístico, palabra o sonido tiene más de un significado. Habitualmente el contexto de la conversación o lectura donde está insertada la palabra deshace los problemas de ambigüedad, pero una palabra polisémica introducida en un sistema documental, sin el contexto, puede dar lugar a ruido documental.

Ejemplo

Un usuario puede estar buscando sobre columnas en arquitectura y recuperar datos sobre columnas tipográficas de diarios. Los lenguajes documentales controlan la polisemia diferenciando cada significado con paréntesis, usando el plural o el singular, adjetivando, etc.

Un tipo de polisemia es la homonimia. La diferencia entre ellas radica en la etimología de la palabra. Si la etimología de las dos palabras es la misma, hablamos de polisemia; si la etimología es diferente, hablamos de homonimia.

Ejemplos de polisemia y homonimia

Misma etimología = polisemia

La polisemia se da cuando una palabra tiene un único origen etimológico y acaba teniendo significados diferentes sin cambiar su categoría gramatical: por ejemplo, no pasa de sustantivo a verbo, como pasa en castellano entre el vino (bebida) y el vino (verbo venir). Es una palabra que con el tiempo ha ido adquiriendo diferentes significados, pero aun así, todos guardan entre sí una relación de significado; por ejemplo, en catalán y castellano fulla/hoja, que viene del latín *folia*, tiene diversos significados, como hoja de una planta, hoja de metal de una herramienta, página de un libro, cada una de las partes de una puerta doble o ventana, etc. Y en todos los significados lleva implícita la idea de una lámina.

Si queremos saber si una palabra es gramaticalmente polisémico, basta con consultar un diccionario etimológico y ver si proviene de un mismo origen. Encontraremos la palabra, un único origen y una lista de diferentes significados. En castellano podemos consultar el *Diccionario de la Real Academia*.

Más ejemplos de polisemia:

- *Servicio*, del latín *servitium*, que ha dado lugar a oficios religiosos, lavabos, misiones militares, cubiertos para comer y, en deportes, poner la pelota en juego. Y en todos ellos permanece la idea de ser útil.

- *Crucero*, del latín *crux*, significando 'cruz', intersección entre las dos naves de una iglesia, encargado de llevar la cruz a la cabeza de una procesión, viaje de placer por el mar, etc. En estos significados la idea es la de la forma de cruz, el cruzar como ir de un extremo a otro.

- *Columna*, del latín *columna*, que usamos para referirnos a los pilares arquitectónicos, las partes verticales de una página impresa de un diario, en física la forma que adoptan algunos fluidos, como "columnas de humo", en el ámbito militar, la formación de barcos o soldados. Y la idea que permanece es la de verticalidad.

Diferente etimología = homonimia

La homonimia se da cuando dos conceptos han llegado a tener el mismo nombre, la misma forma, pero vienen de orígenes diferentes y, por lo tanto, tienen etimologías diferentes.

Por ejemplo, *metro* puede ser el transporte urbano, una unidad de medida o el utensilio para medir. Pero el origen etimológico entre el transporte y los otros dos significados es evidente: el primero es una abreviación de la palabra inglesa *metropolitan*, y en el segundo caso viene del griego μέτρον y significa medida.

Otro ejemplo: la palabra castellana *botín* puede venir del latín *bota* y significará 'calzado hasta el tobillo', o puede venir del alemán *bytin* y significará 'premio de una conquista'.

En castellano y catalán este fenómeno es menos frecuente que en otras lenguas, como el inglés o el francés, en las que abundan las palabras homónimas que dan mucho juego en los chistes.

Dentro de la homonimia podemos diferenciar las palabras que escribiéndose igual tienen significados diferentes, llamadas homógrafas, como las anteriores *metro* o *botín*, de las palabras que sonando igual también tienen significados diferentes, conocidas como palabras homófonas: *vell/bell* en catalán, o *tubo/tuvo* en castellano.

En resumidas cuentas, la sinonimia provoca silencio documental y la polisemia y variantes provocan ruido documental. El control terminológico del vocabulario garantiza el criterio de univocidad que tienen que tener los lenguajes documentales controlados, según el cual un concepto se representa con un término y un término sólo puede tener un significado.

2.1.4. Relaciones de significado de los términos

Por **relaciones de significado** entendemos la relación de genérico, específico o relacionado que puede tener un término con respecto a otro.

En el lenguaje natural estas relaciones son implícitas. Por ejemplo, cuando hablamos de manzanas todos entendemos que se trata de una fruta fresca y que las Fuji y las Golden son variedades concretas. Es decir, situamos el término "manzana" dentro de una jerarquía de términos conceptualmente más genéricos (fruta) y más específicos (Golden, Fuji). Incluso podemos relacionar por

asociación de ideas la manzana con otras frutas, como la naranja o el plátano. Pero en un lenguaje documental hay que definir estas relaciones, agrupando y relacionando los términos afines.

La estructura que relaciona los términos es implícita en el lenguaje natural, pero en los lenguajes documentales hay que hacerla explícita. Eso se puede hacer de dos maneras:

a) En una secuencia jerárquica, donde la propia posición del concepto ya define sus términos genéricos y específicos. También deshace problemas de significado.

Ejemplo de la pesca

Ved el ejemplo de la pesca extraído de la *Clasificación Decimal Universal* (CDU). El concepto *pesca* puede ser la actividad económica o la pesca como deporte. Si nos fijamos en la cadena jerárquica vemos que cada uno cuelga de una clase diferente:

```
6 Ciencias aplicadas. Medicina. Tecnología
63 Agricultura y ciencias relacionadas
 639 Caza. Pesca

7 Bellas artes. Juegos. Deportes
79 Diversiones. Espectáculos. Juegos
 799 Caza deportiva. Pesca deportiva.
```

b) En una presentación alfabética donde cada término se acompaña de todos sus términos relacionados, ya sean equivalentes, genéricos, específicos o relacionados.

El tesauro del CSIC

En el tesauro de Psicología del CSIC[6] , consultamos "Sueños" y encontramos:

Sueños

TG Dinámica de la personalidad

TE Contenido del sueño
TE Pesadilla

TR Déjà vu
TR Interpretación de los sueños
TR Sueño fisiológico
TR Sueño REM
TR Trastornos de consciencia

Las siglas nos informan del tipo de relación que establecen: TG significa término genérico (por encima de "Sueños" el tesauro tiene "Dinàmica de la personalidad"), TE son los términos específicos (son términos específicos de "Sueños": Contenido del sueño, Pesadilla) y los TR son los términos relacionados (se relacionan con "Sueño", "Déja vu", la "Interpretación de los sueños", el Sueño REM", etc.).

(6)Centro Superior de Investigaciones Científicas

Finalmente, las principales ventajas e inconvenientes del lenguaje natural y el documental controlado son:

Ventajas e inconvenientes de los lenguajes documentales

	Ventajas	Inconvenientes
Lenguaje natural	Amigable Actualizado Económico	Dificulta la búsqueda Poco preciso
Lenguaje documental controlado	Unívoco Facilita la búsqueda	Caro Poco actualizado

A modo de conclusión

Indizar es la acción de describir o identificar un documento en relación con su contenido.

La indización la puede realizar una persona (de forma centralizada o de forma coordinada) o un programa.

Por lenguaje natural entendemos el lenguaje que usamos de forma cotidiana (catalán, castellano, vasco), y por lenguaje documental entendemos el listado o vocabulario de términos que usamos para indizar y que puede estar en formato libre o controlado. La principal diferencia entre el lenguaje natural y el documental controlado es el control terminológico:

- El control del número de términos del vocabulario: los lenguajes documentales son entrópicos, tienden a la selección, a la restricción del vocabulario.
- El control de las formas: los lenguajes controlados, controlan las formas plural/singular, el uso de acrónimos y siglas y la construcción de las frases.
- El control del significado: los lenguajes controlados controlan la sinonimia y la polisemia. Decimos que las palabras son sinónimas cuando tienen el mismo significado. Decimos que dos palabras son polisémicas cuando el mismo signo lingüístico tiene más de un significado. La sinonimia provoca silencio documental y la polisemia y variantes provocan ruido documental. El control terminológico del vocabulario garantiza el criterio de univocidad que tienen que tener los lenguajes documentales controlados, según el cual un concepto se representa con un término y un término sólo puede tener un significado.
- Las relaciones de significado entre los términos son las relaciones de genérico, específico o relacionado que puede tener un término con respecto a otro. En el lenguaje natural estas relaciones son implícitas pero en los lenguajes documentales hay que hacerlas explícitas a través de una secuencia jerárquica o una presentación alfabética.

2.2. Cómo se indiza

Ahora que ya hemos visto la necesidad de contar con lenguajes documentales para paliar la ambigüedad del lenguaje natural, estamos en condiciones de preguntarnos por el proceso de indización que lleva a cabo un analista.

A continuación presentamos las **fases** que proponen diversos autores antes de llegar a la que nos servirá como marco de referencia en este subapartado:

- Dos fases: análisis del texto y traducción (Chaumier, 1988; Fidel, 1994).
- Tres fases: análisis del texto, identificación de conceptos y traducción (Amat, 1989; Norma UNE 50-121-91).
- Cuatro fases: análisis del texto, identificación de conceptos, traducción y establecer enlaces sintácticos entre descriptores (Slype, 1991).

- Cinco fases: registro de datos, análisis del texto, identificación de conceptos, traducción y examen de la indización.

En este módulo seguiremos la **norma UNE 50-121-91** y sus tres etapas:

Norma UNE 50-121-91

UNE 50-121-91. *Métodos para el análisis de documentos, determinación de su contenido y selección de términos de indización.*

1) Examinar el documento para identificar su contenido.

2) Seleccionar los conceptos principales del contenido.

3) Traducir a un lenguaje documental.

Ejemplo

Examinamos un libro titulado *Mitos de antiguas civilizaciones*. Leemos el título, el resumen, el sumario, etc.

En una segunda etapa seleccionamos como conceptos principales: Mitos, Grecia, Roma, India, Japón, Indios norteamericanos.

En la tercera etapa indizamos. Si indizamos con un lenguaje libre podemos escribir el término como deseamos o como salga en el texto. Por ejemplo:

```
Mitología india americana.
```

En cambio, si indizamos con un lenguaje controlado tendremos que traducir estos conceptos a una forma controlada. Pongamos por ejemplo que pensábamos indizar Mitología india americana. Veamos cómo quedaría en tres lenguajes documentales diferentes:

```
CDU259.2
LEMACMitologia amerίndia
LEM del CSIC Indios de América - Religión y mitología
```

A continuación se detalla cada parte del proceso.

1) Examen del documento e identificación de los conceptos

El analista tiene que examinar con precisión el documento. La lectura completa es, a menudo, impracticable, pero sí que tiene que prestar atención al título, resumen, sumario, introducción, ilustraciones y palabras o frases destacadas en una tipografía diferente.

No se recomienda la indización sólo a partir del título, ya que hay títulos que llevan a error, y tampoco confiar en que el resumen sea un sustituto del texto, ya que no todos los resúmenes están bien elaborados.

Ejemplo de títulos y resúmenes que no aportan datos significativos para la indización

- Chesneaux, Jean. *¿Hacemos tabla rasa del pasado?* México: Siglo XXI Editores 1981. Su materia es *Historia, historiadores, historiografía*. En el catálogo de la Biblioteca Nacional de España (BNE[7]) lo encontramos indizado como Historia.

- Mallol, Tomas. *Si la memòria no em falla*. Girona: CCG Ediciones 2005.
  Su materia es *Memorias, cine, coleccionismo*. En la Biblioteca de Catalunya (BC[8]) lo encontramos indizado como Cine amateur.

Si recordamos el resumen del libro de Carl Sagan, *Cosmos*, nos daremos cuenta de que no era suficiente para indizar el contenido de la obra. Por estos motivos se recomienda una lectura ágil del resto de partes significativas del documento.

(7)BNE es la sigla de *Biblioteca Nacional de España*.

(8)BC es la sigla de *Biblioteca de Catalunya*.

2) Selección de los términos de indización

Tal como dice la norma UNE, el analista tiene que identificar las nociones que son elementos esenciales de la descripción del contenido. Si la indización es compartida, la institución que la patrocina tiene que establecer claramente los factores que considera importantes.

Para seleccionar los conceptos del documento, el analista tiene que ser consciente del número de conceptos (criterio de exhaustividad) y de la exactitud de los mismos (criterio de especificidad).

a) Exhaustividad

A medida que el analista va leyendo, tiene que ir tomando nota de los conceptos interesantes del documento.

Una buena praxis es la que identifica los conceptos relevantes sobre:

- El tema.
- Los nombres personales que puedan ser interesantes de indizar.
- Los nombres geográficos.
- Las fechas cronológicas.
- La forma en que se presenta el documento: artículo, estadística, formulario o divulgación, científico, etc.

La exhaustividad es un criterio relacionado con el número de conceptos que se tienen en cuenta para caracterizar el contenido entero de un documento. El principal criterio de selección es el valor potencial del concepto para los usuarios de su SID.

Podemos distinguir entre una exhaustividad baja, media y alta en función del número de descriptores. Es en este entorno donde la norma UNE 50-121-91 da sus recomendaciones en cuanto a la exhaustividad. Los criterios que el indizador tiene que tener en cuenta son:

- El tipo de SID y perfil de usuario. No es lo mismo indizar para una base de datos genérica que para una específica.
- El tipo de documento. No se indiza con el mismo número de descriptores una monografía que un artículo de revista, una tesis, etc.

Tal como recomienda la norma UNE, no es conveniente ser estrictos con el número de términos, no se tiene que limitar el número de forma arbitraria, tipo “para una monografía dos términos de indización”, ya que puede conducir a una pérdida de objetividad y a una deformación de la información. Es preferible sugerir un baremo, entre tantos y tantos términos para cada tipo documental y SID y ser flexibles, ya que los criterios que tienen que regir son el propio contenido del documento y su posterior recuperación.

A partir del siguiente resumen informativo, elaboraremos tres tipos de indizaciones sugiriendo un baremo (para esta asignatura y sus prácticas) y una finalidad:

“Anàlisis y descripción de los errores más frecuentes que cometen los profesionales y aficionados a la fotografía astronómica mientras intentan descubrir nuevos objetos celestes todavía no identificados.

Estos errores son debidos a cuatro causas: errores en el proceso de positivado de la copia como consecuencia de la presencia de partículas de polvo en los negativos o en las lentes del equipo de laboratorio; errores en el negativo debidos a defectos de lavado, deficiencias en la emulsión, rayas y rasguños o por el uso de películas de color destinadas a ser forzadas, y errores en las lentes de los objetivos, debidos a efectos de distorsión y a alteraciones en la refracción. Finalmente se describen otras causas: reflejos de la luz del sol sobre las antenas de satélites artificiales Iridum, retoques digitales o de fotocopiadoras y duplicadoras, uso de objetivos sencillos y poco potentes para captar imágenes de cielo profundo y, en último término, oscilaciones del condensador de luz del microscopio.

Todos estos errores pueden dar lugar a imágenes falseadas: objetos inéditos, diámetros erróneos, efectos de redondeo, alineaciones planetarias erróneas, etc. El artículo facilita imágenes de estos errores fotográficos.

Los autores concluyen que hace falta ser cauteloso y hacer las oportunas comprobaciones antes de dar a conocer el descubrimiento de un nuevo objeto celeste a las sociedades astronómicas.”

Cuervo Herrero, C.; Fernández González, A.: “Objetos celestes erróneos”. *Tribuna de Astronomía y Universo. Revista de Astronomía, Astrofísica y Ciencias del espacio*. 2000. II Época, n° 16 – octubre. p. 36-40.

Ejemplo de los tres grados de exhaustividad

Exhaustividad baja	**Exhaustividad media**	**Exhaustividad alta**
Baremo 1-3	**Baremo 4-6**	**Baremo 7...**
Ejemplo de uso: catálogo de una biblioteca pública	Ejemplo de uso: bases de datos de una biblioteca especializada en astronomía	Ejemplo de uso: bases de datos de una biblioteca especializada en astrofotografía

Exhaustividad baja	Exhaustividad media	Exhaustividad alta
Baremo 1-3	**Baremo 4-6**	**Baremo 7...**
Errores fotográficos Fotografía astronómica	Astrofotografía Errores fotográficos Descubrimientos Identificación de objetos celestes Objetos erróneos	Alineaciones planetarias Defectos de lavado Deficiencias de la emulsión Diámetros erróneos Efectos de redondeo Errores en el negativo Errores en el positivado Errores en las lentes Objetos inéditos Objetivos Oscilaciones del microscopio Partículas de polvo Rayadas Reflejos del sol Retoques digitales

b) Especificidad

La especificidad está relacionada con la exactitud en que un concepto particular que aparece en un documento está representado por un término de indización.

Si en el texto que estamos indizando aparece el concepto *Diplomacia*, y este término aparece en el lenguaje documental controlado, tenemos que indizar "Diplomacia". Si indizamos "Relaciones internacionales" o "Embajadores" no estaremos siendo específicos, como podéis ver en la tabla siguiente:

Ejemplo de especificidad

Materia	Correcto, y por lo tanto:	Incorrecto por:	
	Específico	Genérico	Demasiado específico
Diplomacia	**Diplomacia**	Relaciones internacionales	Embajadores

Los conceptos se tienen que identificar de la manera más específica posible, pero en determinados casos se pueden preferir nociones más genéricas:

- Cuando el indizador considere que un exceso de especificidad puede ser negativa en la recuperación; por ejemplo, puede decidir que un modelo muy específico de una máquina se indice con el nombre más genérico de este tipo de máquinas.

- Cuando la idea no esté plenamente desarrollada en el documento, o sólo se haga alusión a ella.

- Cuando se esté a la espera de validar el término más específico.

3) Traducción a un lenguaje documental controlado

Para traducir el concepto inicial escrito en lenguaje natural a un lenguaje documental, el indizador tiene que consultar las listas del lenguaje buscando la forma correcta de introducir el concepto.

Ejemplos

Concepto tal como sale en el texto	Traducción	Lenguaje documental utilizado
Tragicomèdia	791.221.28	Classificación Decimal Universal (CDU)
Eolític	Edat de la pedra	Lista de encabezamientos de materia en catalán
Matriz	Útero	Lista de encabezamientos del CSIC
Monarquía absoluta	Absolutismo	Tesauro de Historia contemporánea del CSIC

Cuando el analista procede a traducir el concepto del texto se puede encontrar en las siguientes situaciones:

a) Encuentra el concepto, solo o repartido por las tablas:

- Consulta el lenguaje y encuentra el concepto a la primera. Entonces indiza con este término de indización. Por ejemplo, buscaba "Eolític" y encuentra que tiene que indizar "Absolutismo".

- Consulta el lenguaje y encuentra el concepto o las partes del concepto repartidos por el lenguaje. Entonces tiene que conocer las reglas de combinación de las partes integrantes del término de indización. Ejemplos:
  - Una notación con CDU como 391.91(961.3) "Tatuajes de la isla de Samoa" está formada por 2 elementos, tatuajes + Samoa. Estos elementos van colocados en un orden determinado por las reglas de precoordinación de la CDU (primero la clase principal + auxiliar).

  - Un encabezamiento construido con la LEM del CSIC como Agua-Aspectos económicos está formado por dos partes: Agua + Aspectos económicos, que es un encabezamiento y un subencabezamiento respectivamente y van en este orden.

Con los lenguajes tesauros y listado de autoridades no hay una sintaxis de combinación.

b) No encuentra el concepto:

- Consulta el lenguaje y no encuentra el concepto. Entonces el indizador tiene que conocer las obras de referencia que su SID considera como autoridades reconocidas en la materia. Estas obras de referencia son diccionarios, enciclopedias, otros lenguajes documentales (especialmente los tesauros construidos de acuerdo con las normas ISO y UNE 50-106 y UNE 50-125), atlas, etc.

- Hay lenguajes, como tesauros, donde el indizador tiene que proponer el término nuevo como descriptor candidato y esperar a que la dirección del tesauro lo valide como descriptor. Mientras tanto indiza con un término más genérico.

2.3. Lenguajes documentales

Para indizar necesitamos los lenguajes documentales, que son vocabularios de términos que facilitan la representación del contenido de los documentos.

> Las principales funciones de los lenguajes documentales son indizar el contenido de los documentos y permitir su recuperación a partir del campo materia.

Tercera función de los lenguajes documentales

Existe una tercera finalidad, que solo se da en los sistemas de clasificación: la ordenación altamente significativa del fondo documental del SID.

Los lenguajes documentales son de seis tipos:

1) los sistemas de clasificación,
2) las listas de encabezamientos de materia,
3) las listas de autoridades,
4) los tesauros,
5) las listas de descriptores libres, y
6) las listas de palabras clave o indización automática.

Los términos de indización

Cada lenguaje documental proporciona un nombre diferente a su término de indización y es conveniente que, cuando nos expresemos, lo hagamos con propiedad.

Términos de indización

Lenguaje documental	Su término de *indización* se conoce como	Ejemplo
Sistemas de clasificación	Notación o símbolo de clase	351.851:069 (Ley de Museos)
Listas de encabezamientos de materia	Encabezamiento	Francés-argot
Listas de autoridades	Autoridad, identificador o descriptor	Bécquer, Gustavo Adolfo, 1836-1870
Tesauro	Descriptor	Ramon Berenguer III el Gran NA: [1097-1131]
Listas de descriptores libres	Descriptor	Semana_santa
Listas de palabras clave	Palabra clave	Metro

Existe otro término, denominado **unitérmino**, que no hace referencia a ningún lenguaje documental concreto, sino al hecho de que el término de indización sea simple o compuesto.

> La Norma UNE 50-113-92/1 define los unitérminos como el elemento significativo más pequeño de un lenguaje documental utilizado para representar un concepto específico en un sistema de indización coordinado; no se tiene que confundir con palabra clave o descriptor.

El descriptor *Semana Santa* está formado por dos unitérminos: *Semana* y*Santa*. Y el descriptor *Navidad* está formado por un único unitérmino.

Diferencia entre descriptor y unitérmino

Una palabra	Más de una
Navidad	Semana Santa

Hay que prestar atención al término **palabra clave** porque su uso en la bibliografía científica tiene varias aplicaciones que nos pueden confundir. Es habitual encontrar en los artículos un apartado, bajo el resumen, denominado "palabras clave", en el que el autor nos da los términos que considera más representativos del texto. Estas palabras clave son muy a menudo descriptores de procedencia desconocida (no sabemos si son libres o controlados). En cambio, en este material docente, *palabra clave* se entiende como el término de indización proveniente de la indización automática habitualmente coincidente con un unitérmino.

Las tipologías de los lenguajes documentales

Las tipologías de los lenguaje documentales son los criterios que nos permiten agrupar o clasificar los seis lenguajes documentales en categorías afines. Son las siguientes:

Reflexión

Si domináis las tipologías, podréis responder a cuestiones del tipo: comparad lenguajes, buscad ventajas e inconvenientes, causas de la complementariedad, etc. Se recomienda que las interioricéis.

1) Naturaleza: codificado o natural

Por **codificado** entendemos el uso de un código artificial compuesto por números, letras y símbolos que traducen un concepto. Solo existe un tipo de lenguaje codificado: los sistemas de clasificación.

Ejemplos de términos de indización codificados

CDU	DDC	LCC
94	483	RE 1-994

Por **natural** entendemos el uso de palabras del lenguaje usual, habitual, no códigos. Es mucho más próximo al usuario, más amigable. Hay cinco lenguajes documentales naturales: las listas de encabezamientos de materia, las listas de autoridades, los tesauros, las listas de descriptores libres y las listas de palabras clave.

Siguiendo el ejemplo anterior:

Ejemplos de términos de indización naturales

Historia	Diccionarios de griego clásico	Oftalmología

2) Control: libre o controlado

Un vocabulario libre es una lista de términos extraídos del lenguaje natural sin sufrir ningún tipo de actuación sobre el número de términos, la forma (singular, plural, masculino, femenino), el significado (sinónimo, polisémico) o las relaciones entre los términos.

Normalmente, los lenguajes libres se usan en sistemas automatizados en los que hay un fichero inverso o diccionario de la base de datos. Presentan numerosas ventajas en la indización, como por ejemplo el gasto mínimo de construcción, la actualización inmediata, una máxima coherencia y la riqueza terminológica. Sin embargo, plantean inconvenientes en la recuperación, ya que, al trabajar con lenguaje natural, arrastra todos los problemas derivados de la ambigüedad (sinonimia, polisemia, homonimia). Hay dos tipos de lenguajes libres: las listas de descriptores libres y la lista de palabras clave.

Un **vocabulario controlado** es una lista previamente redactada de términos que se consideran aceptados y unívocos para la indización. Solo los términos de la lista se pueden emplear para indizar.

Se trata de términos seleccionados tanto en su forma (plural, singular, sintagma nominal, adjetivo, siglas, etc.) y en su contenido (se elige un sinónimo de todos los posibles, los homónimos se diferencian entre ellos con paréntesis o adjetivos, etc.) como en sus relaciones de jerarquía y asociación (términos conceptualmente más genéricos o específicos y términos que se evocan mutuamente). Requieren unos gastos de construcción elevados, no solo en personal cualificado, sino también en tiempo. Para muchos autores, son los verdaderos lenguajes documentales. También se conocen con el nombre de **lenguajes artificiales**.

Su función documental es la de representar un concepto con un único término y que solo haya un término por concepto, lo que se conoce como **univocidad**.

Los lenguajes controlados son cuatro:

- los sistemas de clasificación,
- las listas de encabezamientos,
- las listas de autoridades, y
- los tesauros.

Ejemplos de términos libres y controlados

Concepto	Libre	Controlado
Limpieza	Higiene, Limpieza, Profilaxis, Aseo, Sanidad, Desinfección	CDU: 613 LEMAC: Higiene

3) Coordinación: precoordinación o poscoordinación

La **precoordinación** consiste en determinar a priori cómo se combinan los términos, tanto en la construcción del lenguaje como a la hora de indizar o recuperar el documento.

La precoordinación en las bibliotecas manuales

La precoordinación era una auténtica necesidad en el entorno de las bibliotecas manuales (fichas de cartulina), ya que no se podía buscar por una combinación de dos términos o más.

Asimismo, se hace referencia a la precoordinación como la sintaxis del lenguaje documental. Por ejemplo, en las listas de encabezamientos de materia, los epígrafes siguen un orden concreto para evitar la dispersión de encabezamientos.

Así, un documento de congresos catalanes sobre arqueología submarina se indizaría como Arqueología submarina – Catalunya – Congresos, y no con ninguna otra de las **combinaciones posibles**.

Combinaciones posibles

Las combinaciones erróneas son las siguientes:

- Catalunya – Congresos – Arqueología submarina
- Arqueología submarina – Congresos – Catalunya
- Congresos – Arqueología submarina – Catalunya
- Arqueología submarina – Congresos – Catalunya

Recordemos que el orden viene determinado por las indicaciones que acompañan a cada epígrafe. Así, vemos que *Arqueología submarina* puede llevar subdivisión geográfica y que *Congresos* es una subdivisión que puede ir detrás de nombres propios de persona, familias, entidades, clases de personas, grupos étnicos, guerras y temas; por lo tanto, el único orden posible es el de la solución aportada.

Existen dos lenguajes precoordinados: los sistemas de clasificación y las listas de encabezamientos de materia.

La **poscoordinación** consiste en indizar términos sueltos. No tienen sintaxis en el momento de la indización, sino que se combinarán a la hora de la recuperación siguiendo la lógica de los operadores booleanos.

Cada término indizado es un punto de acceso al documento: cuanto más términos indicemos, mayor es la posibilidad de recuperarlo. Siguiendo con el caso anterior, lo formularíamos poniendo los tres conceptos en cualquier orden, ya que no resulta relevante, por ejemplo:

Congresos and Catalunya and Arqueología submarina

Existen cuatro lenguajes poscoordinados: las listas de autoridades, los tesauros, las listas de descriptores libres y la indización automática.

4) Estructura: jerárquica o alfabética (combinatoria)

En la **estructura jerárquica** o sistemática, el vocabulario se presenta en forma de arborescencia, con términos genéricos que agrupan otros más específicos. Todos los términos dependen de un término superior y de significado más genérico. Esta estructura permite agrupar los conceptos por temas, así como situarlos en su contexto, ya que la secuencia jerárquica nos informa del campo temático al que se adscribe el concepto.

Ejemplo

Pongamos como ejemplo el concepto *libertad*, que tiene muchas acepciones. Simplemente viendo dónde está insertado, ya deducimos si se trata de la libertad filosófica, de derechos humanos o de la libertad de movimientos en máquinas.

La estructura jerárquica informa del campo del conocimiento.

Clase 1	Clase 3	Clase 6
123 Libertad y necesidad 123.1 LIBERTAD. INDETERMINISMO 123.11 Casualidad 123.2 NECESIDAD 123.21 Fatalismo	342.7 DERECHOS FUNDAMENTALES. DERECHOS HUMANOS. DERECHOS Y DEBERES DE LOS CIUDADANOS 342.71 Nacionalidad. Ciudadanía 342.72/.73 Derechos de los ciudadanos. Derechos civiles. El Estado y el ciudadano 342.721 Libertad individual. *Habeas corpus*	62-23 ENGRANAJES. ELEMENTOS MECÁNICOS DE TRANSMISIÓN. DISPOSITIVOS TRANSPORTADORES Y DE SUJECIÓN 62-231 Estructuras de los mecanismos de transmisión 62-231.2 Sistemas lineales. Pares cinemáticos 62-231.21 Sistemas sin grados de libertad. Acoplamiento automático. Centrado automático 62-231.22 Sistemas con un grado de libertad. Cojinete. Barra de guía. Par de roscado (tornillo y tuerca)

Los lenguajes jerárquicos son dos: los **sistemas de clasificación** y los **tesauros** (en la parte de presentación sistemática o jerárquica).

En la **estructura combinatoria**, los términos no forman cadena, sino que se organizan en listas por orden alfabético. Este tipo de estructura surgió como contrapunto a la rigidez de la estructura jerárquica, que no era fácil de actualizar.

Ejemplo extraído de la *Lista de encabezamientos* del CSIC.

Árbol de la papaya

Árbol de la vida

Árbol del conocimiento

Árboles

Árboles – Crecimiento

Árboles – Cuidados

Árboles – Cultivo

Árboles – Culto

La estructura combinatoria permite la inclusión de términos nuevos y la eliminación de los obsoletos sin que esto afecte al resto de la estructura del lenguaje.

En la secuencia anterior podríamos incluir: Árboles – Adobo, sin alterar el resto.

La facilidad para actualizar el vocabulario los convierte en lenguajes adecuados para todo tipo de entornos: enciclopédicos, científicos y técnicos. Los lenguajes de estructura combinatoria son cinco:

- las listas de encabezamientos de materia,
- las listas de autoridades,
- los tesauros,
- la lista de descriptores libres, y
- las listas de palabras clave.

Tesauro

Como podéis observar, el tesauro participa de las dos estructuras: tiene una presentación sistemática en forma jerárquica y una presentación alfabética en forma combinatoria.

5) Análisis: por materias, por conceptos o por palabras clave

La diferencia entre uno y los otros estriba en indizar un tema del documento, varios conceptos o todas las palabras con significado.

Reflexión

Hoy en día, la evolución y automatización de los sistemas de información posibilitan que estos lenguajes, en origen sintéticos, puedan indizar de manera más analítica, en especial los encabezamientos de materia, que pueden indizar dos, tres o cuatro encabezamientos. O las notaciones con sistemas de clasificación, que duplican el campo 080 del MARC.

a) Por materias

Es la indización más sintética: indiza uno o dos términos de indización. Responde a la pregunta "¿cuál es el tema de este documento?". Existen dos lenguajes que indizan por materias: los sistemas de clasificación y las listas de encabezamientos de materia.

b) Por conceptos

Responden a la pregunta "¿cuáles son los conceptos de este documento?". Van ligados necesariamente a sistemas automatizados, ya que no sería factible elaborar tantas fichas de cartulina como conceptos se indizaran. Existen tres lenguajes que indizan por conceptos: las listas de autoridades, los tesauros y las listas de descriptores libres.

c) Por palabras clave

Indizar por palabras clave representa indizar todas y cada una de las palabras con significado del texto. Es el proceso más analítico que hay. No se trata de una tarea de indización humana, sino automática. Solo hay un lenguaje por palabras clave, y es evidentemente el único lenguaje automático: la lista de palabras clave.

Resumen de las tipologías

		Sistemas de clasificación	Listas de encabezamientos de materia	Listas de autoridades	Tesauros	Lista de descriptores libres	Lista de palabras clave
Según la naturaleza de los términos	Codificado	X					
	Natural		X	X	X	X	X

		Sistemas de clasificación	Listas de encabezamientos de materia	Listas de autoridades	Tesauros	Lista de descriptores libres	Lista de palabras clave
Según el nivel de control sobre los términos	Libre					X	X
	Controlado	X	X	X	X		
Según el nivel de coordinación de los términos	Precoordinado	X	X				
	Poscoordinado			X	X	X	X
Según la forma de agrupar los términos o estructura	Jerárquico	X			X		
	Alfabético		X	X	X	X	X
Según el nivel de análisis	Por materias	X	X				
	Por conceptos			X	X	X	
	Por palabras clave						X

Una buena praxis es estudiar los seis lenguajes según la tipología y recordar fórmulas como por ejemplo:

1 codificado + 5 naturales = 6

4 controlados + 2 libres = 6

2 precoordinados + 4 poscoordinados = 6

2 jerárquicos + 4 combinatorios = 6

2 por materias + 3 por conceptos + 1 por palabras clave = 6

2.3.1. Clasificar y recuperar con sistemas de clasificación

Este apartado apuesta por redescubrir la potencia combinatoria de los sistemas de clasificación y comprobar su estado actual. Constataremos que, si bien son muy prácticos en la indización, no lo son tanto en la recuperación en línea, al menos por el momento.

Sistemas de clasificación en la Web

De los nueve principales sistemas de clasificación implementados en estos momentos en todo el mundo, seleccionamos tres para hacer las prácticas de este módulo, aunque el porcentaje más elevado de prácticas lo haremos con la clasificación decimal universal, en la versión abreviada en español:

Sistemas de clasificación documental vigentes

Los sistemas de clasificación documental vigentes son los siguientes: clasificación decimal universal (CDU), clasificación Dewey (DDC), clasificación de la Library of Congres (LCC), clasificación china, clasificación japonesa, clasificación rusa (LBC, antigua BBK), clasificación Colon (CC), clasificación Bliss (CB) y clasificación Brown.

1) Clasificación decimal universal (CDU)

- Universal Decimal Classification Consortium Homepage (2002, 1 de agosto) [en línea]. La Haia: UDC Consortium. Act. 2002-08-01. [Fecha de consulta: 10 de octubre del 2008.]

2) Clasificación decimal Dewey (DDC)

- http://www.oclc.org/dewey/resources/summaries/default.htm, 025.431: The Dewey blog [en línea]. [Fecha de consulta: 10 de octubre del 2008.]

- Online Computer Library Center. Dewey services, Dewey decimal classification for use with OCLC's online cataloging services [en línea]. [Fecha de consulta: 10 de octubre del 2008.]

3) Clasificación de la Library of Congress (LCC)

- Library of Congress Classification system [en línea]. [Fecha de consulta: 1 de octubre del 2008.]

Clasificación en la actualidad

Los sistemas de clasificación son más que centenarios. Están considerados los primeros lenguajes documentales verdaderos y, desde su generalización en las bibliotecas en el siglo XIX, han demostrado su eficacia recuperando por materias. Ahora bien, no han estado exentos de los embates de la crítica, ya que algunas de sus características inherentes (como el tiempo que requieren, la síntesis o la codificación) no parecían encajar en momentos de explosión documental, de acceso a grandes bases de datos y en red.

Sistemas de clasificación

Los sistemas de clasificación son lenguajes controlados, codificados, precoordinados, sistemáticos o jerárquicos y sintéticos por materias.

La década de 1960 supuso un momento crítico, al cuestionarse que los sistemas de clasificación fueran el lenguaje documental adecuado para abarcar la gran cantidad de documentación científica que se iba generando (documentación cada vez más abundante y, por lo tanto, lenta de clasificar), con terminología nueva (que la lentitud de las actualizaciones haría imposible de asumir), con necesidades nuevas como acceder por conceptos y palabras (cuando las clasificaciones optaban por materias).

Otro embate, este más reciente, ha sido el papel que pueden tener estos sistemas en un entorno web, donde imperan los paradigmas de la indización social y la indización automática. En este contexto, ¿tienen sentido las jerarquías y las notaciones codificadas?

Afortunadamente, todos los lenguajes documentales tienen cabida en la representación del conocimiento. Las jerarquías, también llamadas presentaciones sistemáticas, arborescencias o incluso taxonomías, presentan una virtud excepcional a la hora de indizar y recuperar, y es que permiten situarnos en una secuencia de términos más genéricos o más específicos; por lo tanto, podemos elegir el grado de especificidad y el término en el contexto que nos interesa.

En la cadena siguiente observamos cómo se abre el concepto *religión* hasta llegar a las religiones específicas del hinduismo. El analista decidirá si indiza con una clase más genérica o más específica. La decisión dependerá de las necesidades del SID. Por ejemplo, un SID especializado en documentación sobre religiones probablemente indizará de manera específica y escogerá uno de los tres últimos.

2 Religión
23 Religiones del subcontinente indio
231 Vedismo
232 Brahmanismo
233 Hinduismo
233.2 Visnuismo
233.3 Shivaísmo 
233.4 Shaktismo

En el ejemplo siguiente observamos que la posición dentro de una cadena nos informa del contexto de cada concepto. Podemos localizar el concepto *iglesia cristiana* en la clase 27 Religión o en la 726.54 Arquitectura, según si nos interesa un enfoque de la fe o de la arquitectura.

Ejemplo de enfoque

2 Religión	**7 Arte**
27 Cristianismo. Iglesias cristianas	72 Arquitectura 726 Arquitectura religiosa 726.5 Arquitectura de las iglesias 726.54 Iglesia

Esta elección es posible en cuadros jerárquicos, no en listas alfabéticas que resuelven el tema de los enfoques reservando el término simple para un tema y creando uno compuesto para el otro.

En la LEMAC se soluciona de la manera siguiente:

Solución en una lista de encabezamientos de materia.

Religión	**Arte**
Iglesia	Arquitectura religiosa

A los **sistemas de clasificación** se les reconoce el papel principal que han tenido a la hora de estructurar el conocimiento creando sistemas que permitían representar y recuperar los datos a partir del significado de los documentos, es decir, a partir de la materia y no de datos formales como nombres propios o títulos.

Las estructuras clasificatorias son elementos muy importantes en la organización del conocimiento. Nos permiten representar y ordenar el conocimiento, y esto, en un momento como el actual, en el que la información está cada vez más atomizada y dispersa, hace que los sistemas de clasificación nos proporcionen una visión coherente y homogénea, una perspectiva integradora.

Por lo que respecta a las **notaciones**, los códigos numéricos o alfanuméricos, ¿todavía suponen una buena opción ante el uso amigable del lenguaje natural? Esta pregunta equivale a interrogarse sobre si un lenguaje documental codificado tiene suficientes utilidades para merecer la inversión en tiempo y esfuerzo. Pues bien, obtendremos la respuesta observando las ventajas que representa la codificación, y que son las siguientes:

- Los códigos son internacionales y, por lo tanto, la codificación permite el intercambio (en red de ámbito nacional o internacional).

- Permite ordenar el fondo y disponerlo en anaqueles de manera altamente significativa.

- Permite elaborar tanto productos bibliográficos como bibliografías nacionales o selectivas (existe constancia de que la CDU se usa al menos en treinta bibliografías nacionales).

- Permite confeccionar índices y guías por materias.

- Permite difundir de forma selectiva la información (DSI).

Observación

En la bibliografía científica encontraréis que contraponen la codificación de las clasificaciones con el lenguaje natural, no con el lenguaje libre, por lo que el principal inconveniente de los sistemas de clasificación no es que sean controlados, sino que están codificados. Si el problema fuera el control, otros lenguajes, como los encabezamientos de materia, las autoridades y los tesauros, también recibirían la misma crítica.

Encuesta sobre el uso de la CDU

En una encuesta del Consorcio de la CDU (Aida Slavic, 2007) llevada a cabo en doscientos siete países del mundo, se concluyó que ciento veinticuatro países (el 60%) clasificaban con CDU. De estos países, treinta y cuatro (el 28%) tienen la CDU como sistema principal, cuarenta y cinco (el 36%) la usan en determinados tipos de bibliotecas y los cuarenta y cinco restantes (el 36%) solo la usan en algunas bibliotecas de sus naciones.

Como hemos visto, las estructuras, las jerarquías y los códigos tienen su utilidad; aun así, los sistemas de clasificación han evolucionado y han mejorado tres aspectos básicos: la **estructura**, el **contenido** y la **visibilidad en la Web**.

Estructura

En el caso de la CDU, que es la clasificación que más trabajaremos, la mejora de la estructura pasa por potenciar la**facetación** (Broughton, 2009).

Las facetas son principios de división, características que las materias tienen en común. Las facetas agrupan los conceptos según una característica concreta que comparten con otras clases.

Hay facetas de tipo **universal**, aplicables a todos los campos del saber (como el tiempo y el espacio), y las**propias de una materia**.

Ejemplo de facetas

El espacio, el tiempo, la forma, la lengua. Por ejemplo, dentro de la faceta *forma*podemos encontrar *miniatura*, que podremos aplicar a todo tipo de conceptos, como:

- diccionarios en miniatura,
- pintura en miniatura,
- modelismo en miniatura.

Ejemplo de facetas propias de una materia

Facetas para la materia Arte: Periodo artístico, Técnica artística, Tema representado... Así, dentro de *tema representado* podemos encontrar *figura humana*, que podríamos aplicar a cualquier tipo de arte, como pueden ser:

- la figura humana en escultura,
- la figura humana en pintura,
- la figura humana en los esmaltes.

Se considera que la CDU es una clasificación mixta o híbrida porque combina una estructura enumerativa con una facetada. Para combinar dispone de dieciséis signos, nueve tablas auxiliares y un número muy elevado de auxiliares especiales repartidos por todas las tablas, principales e incluso auxiliares.

Ventajas e inconvenientes de la facetación

Ventajas de la facetación:

- Es analítica y, por lo tanto, permite describir con precisión el contenido de un documento.
- Es flexible y no queda desfasada con rapidez.
- Es fácilmente automatizable, ya que los documentos se pueden buscar en conjunto o para cada faceta.

Inconvenientes de la facetación:

- Su aplicación es compleja.
- Hay muchas materias que no se pueden representar fácilmente con facetas (conceptos de tipo mental que no son objetos).
- No todos los documentos tienen todas las facetas, lo que hace que la notación no sea homogénea.

Las tablas de auxiliares, tanto comunes como especiales, son pequeñas clasificaciones jerárquicas autónomas a partir de una faceta (la forma, el lugar, el idioma, etc.) que, cuando se combinan, describen la materia del documento de forma analítica. Son tablas autónomas pero articuladas, combinables, que dan más flexibilidad a las tablas enumerativas.

¿Cómo se potencia la facetación?

La CDU ya es una clasificación mixta, así pues, ¿cómo se potencia la facetación? Retirando clases compuestas (antes la clase 29 comprendía las religiones antiguas –hinduismo, judaísmo, islam– y ahora se encuentran asignadas cada una a una clase propia; por ejemplo, la 26 Judaísmo, la 27 Cristianismo y la 28 Islam), revisando radicalmente clases como la informática o la medicina, proponiendo un orden más claro en la construcción y una sintaxis más lógica y eliminando progresivamente el signo subdividir en favor del *colon* (:) o auxiliares especiales.

Contenido

Para mejorar el contenido, el Consorcio de la CDU revisa cada año las tablas y publica cada mes de noviembre los cambios en el documento *Extensions and corrections to the UDC*, y edita electrónicamente el 1 de enero siguiente el *master reference file*, o fichero básico de referencia, en el que comunica las eliminaciones, correcciones y ampliaciones.

En términos generales, las clases que han evolucionado más son las 004 Informática, 2 Religión, 61 Medicina, 8 Lengua y Literatura y 9 Geografía e Historia.

Web recomendada

Podéis consultar los cambios en la web del Consorcio de la CDU. *Major changes to the UDC since 1993* (http://www.udcc.org/major_changes.htm).

Visibilidad

Para mejorar la visibilidad en la Web disponemos de los **metadatos**, que resultan clave en el proceso de captación y transmisión de estos significados y de los estándares para ontologías.

El uso de los metadatos es muy desigual. En el campo materia se puede poner el término de indización en varios lenguajes, entre ellos la CDU (o*UDC* en inglés).

La iniciativa del Dublin Core recomendó en el documento *Dublin Core Qualifiers* del año 2010 los siguientes lenguajes documentales:

- DDC, http://dublincore.org/documents/dcmes-qualifiers/#ve-DDC
- IMT, http://dublincore.org/documents/dcmes-qualifiers/#ve-IMT
- LCC, http://dublincore.org/documents/dcmes-qualifiers/#ve-LCC
- LCSH, http://dublincore.org/documents/dcmes-qualifiers/#ve-LCSH
- MESH, http://dublincore.org/documents/dcmes-qualifiers/#ve-MESH
- NLM, http://dublincore.org/documents/dcmes-qualifiers/#ve-NLM
- TGN, http://dublincore.org/documents/dcmes-qualifiers/#ve-TGN
- UDC. http://dublincore.org/documents/dcmes-qualifiers/#ve-UDC

Webs recomendadas

Dublin Core Metadata Initiative: Metadata Terms http://dublincore.org/documents/2010/10/11/dcmi-terms/

Las siglas corresponden a IMT: Internet Assigned Numbers Authority http://www.iana.org/assignments/media-types/

NLM: National Library of Medicine Classification http://wwwcf.nlm.nih.gov/class/

TGN: Tesaurus Getty de nombres geográficos http://www.getty.edu/research/tools/vocabulary/tgn/index.html

Los metadatos pueden ir asignados al documento o estar separados de este. Los primeros están integrados en el documento, de modo que si el documento cambia de ubicación, los metadatos también lo hacen, y el recopilador lo localiza y lo incluye en el índice. Solo podemos acceder a los metadatos y actualizarlos acudiendo a la misma fuente. Los separados del documento tratan la materia como un objeto de información en sí y van a una base de datos aparte. Podemos acceder a los metadados sin acceder al recurso. Es el tipo de metadato más parecido a los registros bibliográficos.

Los estándares para ontologías, como el **formato SKOS**, son un vehículo para el despliegue de sistemas de organización del conocimiento que no han nacido digitales (o XML/RDF), como los tesauros y las clasificaciones bibliográficas. El consorcio de la CDU, en sus *Machine readable files & linked data* (http://www.udcc.org/udcsummary/exports.htm), lanza el formato SKOS en veintiséis idiomas. Los problemas no resueltos se encuentran en la aplicación y recuperación de las facetas y en la coordinación de los términos con signos y auxiliares (Devika Madalli, 2009).

SKOS

Sistemas de organización del conocimiento (SKOS), un modelo de datos común para compartir y enlazar sistemas de organización del conocimiento mediante la web semántica.

Contextos en los que clasificamos

Actualmente, los sistemas de clasificación son útiles en los contextos que describimos a continuación.

Clasificación de una colección de documentos

La clasificación de una colección de documentos es la aplicación clásica de este lenguaje, a pesar de que Paul Otlet y Henry La Fontaine lo hicieron extensible a la clasificación de la bibliografía universal.

Podemos clasificar el fondo en nuestra base de datos con una **notación completa** o con una**simplificada**. La primera describe con más especificidad el contenido y es más compleja de recuperar, mientras que la segunda es más genérica pero más intuitiva. Ahora bien, la misma estructura decimal de la notación nos facilita la elección, ya que se presta a las dos opciones: un gran número de bibliotecas utiliza una versión simplificada de la CDU para organizar los anaqueles (resulta fácil para los documentalistas y los usuarios a la hora de localizar el documento), pero dentro del catálogo las notaciones tienen un mayor desarrollo, de forma que se lleva a cabo un análisis más específico. Así, tanto el documentalista como el usuario pueden echar un vistazo a los estantes y también buscar en el catálogo de forma más exhaustiva.

Los códigos de clasificación también permiten confeccionar estadísticas sobre el volumen de la colección y las temáticas más consultadas o prestadas.

Ordenación de manera altamente significativa de un fondo documental

Recordemos que existen tres tipos de ordenaciones; se elegirá uno en función de si el acceso a las estanterías es **libre o no**.

- **Ordenaciones no significativas** (también conocidas como numéricas): son aquellas en las que no existe relación con el contenido del documento. Por ejemplo, ordenar según el número asignado a la llegada. Es una ordenación útil en los SID que no den acceso libre a los anaqueles y en los que la recuperación la haga el documentalista. Apropiado para los SID que tengan su colección en el depósito y no en la sala de lectura. Es el método más empleado en archivos.

- **Ordenaciones con significado limitado**: se ordenan por algún criterio como, por ejemplo, la lengua del documento, el autor o el tema. Es la ordenación que encontramos en librerías, en bibliotecas personales o en bibliotecas pequeñas.

- **Ordenaciones altamente significativas**: se ordena a partir de un cuadro de clasificación, de forma que los contenidos afines se colocan uno junto a otro. Es apropiada para los SID de libre acceso como las bibliotecas públicas y universitarias. Un ejemplo sería ordenar siguiendo las clases de la CDU.

Los tres tipos de ordenaciones se pueden combinar en un mismo SID.

En una biblioteca pública se puede ordenar de la manera siguiente:

- Ordenaciones no significativas: prensa, boletines, revistas...
- Ordenaciones con significado limitado: novelas. Se hace distinción entre novela histórica, ciencia-ficción, biografías, etc. Dentro de cada grupo, las novelas se ordenan alfabéticamente por autores.
- Ordenaciones altamente significativas: todo el resto de la colección. Es el grueso más importante.

James Duff Brown

James Duff Brown fue un bibliotecario británico que, mientras era director de la Clerkenwell Public Library de Londres en 1893, organizó por primera vez el acceso libre a las estanterías o anaqueles abiertos.

Según Foskett (1996), existen dos razones para clasificar de manera altamente significativa en estantería abierta:

- la **primera** es satisfacer la curiosidad del usuario, ya que puede recorrer los anaqueles buscando un ítem que le resulte atractivo, y
- la **segunda** es que muchas veces, a partir de un documento, localiza a su alrededor otros ítems interesantes.

Para ordenar con un sistema decimal hay que colocar las notaciones en este orden: 1, 11, 111, 2.

En el caso de combinaciones con signos clasificatorios y auxiliares especiales, el orden es el siguiente:

- Auxiliares comunes independientes = (0...) (1/9) (=...) y "..."
- Firmas + /
- Clase expresada como número simple (por ejemplo, 622.341.1)
- *Colon* (:), doble *colon* (::), auxiliar de lengua, de forma, de lugar, de razas, de tiempo, asterisco, A/Z,.00,-0-1/-9,.0'.
- Número simple siguiente.

Lectura recomendada

Consultad la introducción de la CDU (edición abreviada), página XXVII, donde se ejemplifica cómo se tienen que ordenar todas las posibilidades entre el código 622.341.1 y el código 622.341.11. Muy ilustrativo.

Clasificación de documentos web

La gran cantidad de documentos web que se genera cada día hace imposible su clasificación manual. Ahora bien, clasificar no es solo una operación intelectual, también se puede llevar a cabo de manera automática en algún momento del proceso. Hay tres líneas de trabajo (Moreno, 2002) para clasificar automáticamente la Web.

1) Método de clasificación a priori: método automático que usa una clasificación establecida a priori para asignar notaciones. El programa detecta los conceptos del recurso web expresados en lenguaje natural y los traduce al código de la clasificación. El programa trabaja con índices de materias a modo de listas de autoridades que pueden haber sido generados manualmente o por robots procedentes de la misma clasificación o de otros lenguajes documentales naturales. Incluso es recomendable incrementar la lista con términos no expresados en el cuadro de clasificación pero que los usuarios usan en las consultas.

El robot lee en el título del documento *magnoliófitas*, consulta el fichero de autoridades, ve que es un término sinónimo de *angiospermas* y traduce 582.5/.9. El documento queda indizado con esta notación.

2) La clusterización: se trata de una técnica de clasificación automática también conocida como clasificación derivada o a posteriori, que consiste en agrupar documentos relacionados entre sí por el tema formando conjuntos o clústeres. La diferencia estriba en el hecho de que este último proceso se efectúa automáticamente.

Clasificación de textos científicos o literarios

Es más fácil clasificar automáticamente los textos científicos (terminología más precisa) que los literarios (términos más ambiguos). Un texto sobre cine titulado ***Senderos de gloria*** podría quedar clasificado como 625.711.2, que significa **carreteras, caminos**.

Un ejemplo sería el buscador Yippy http://yippy.com/, que muestra los resultados clasificados en carpetas e indica el número de documentos que contiene cada una. El número y el nombre de las carpetas varía en cada búsqueda, se van creando dinámicamente de manera automática según los resultados que vuelca la base de datos. Así, si hacemos una búsqueda por *library of congress subject headings*, nos aparecerán unas carpetas diferentes que si buscamos *universal decimal classification*, sin coincidir en las carpetas que comparten conceptualmente, que podrían ser *indización*, *lenguaje documental* o *análisis de contenido*.

3) Conversión automática entre sistemas de clasificación:

El proyecto de reclasificación automatizada de la Biblioteca de la Universidad de Kentucky, con el que cambian la CDD por el LCC Library of Congress Classification System. Se trata de mecanizar el proceso de conversión entre dos lenguajes documentales; por ejemplo, de la LCC a la DDC o a la inversa (http://www.questionpoint.org/crs/html/help/it/ask/ask_map_lcctoddc.html).

Recursos web clasificados

En la Web encontramos bases de datos bibliográficas, directorios y portales de información que presentan sus recursos de forma clasificada, pensados para la navegación (*browsing*). El cuadro de clasificación puede ser de elaboración propia o a imagen y semejanza de cuadros sistemáticos clásicos, como la CDU, la DDC o la LCC. Se consideran **productos de información de alto valor añadido**, ya que están elaborados por un equipo humano que identifica el contenido de los recursos (descripción e indización) de manera más cuidadosa que la que hace un robot. Se trata de un tipo de fuente que apuesta por la precisión (documentos seleccionados por la calidad del contenido) frente a la exhaustividad. En esta categoría, además de los directorios, también se incluyen los índices temáticos, las guías temáticas o los *Internet subject gateways*. Sin embargo, hay que añadir que el nivel de clasificación es muy sucinto (uno o dos dígitos a lo sumo).

Webs recomendadas

En *The role of classification schemes in internet resource description and discovery*, de www.ukoln.ac.uk, en el año 2001 se contabilizaron treinta y cinco sistemas de clasificación diferentes usados en distintos portales y directorios temáticos.

Y en *Beyond Bookmarks: Schemes for Organizing the Web*, http://www.public.iastate.edu/CYBERSTACKS/CTW.htm, encontraréis una lista de las bases de datos organizadas según el sistema de clasificación que usan, ya sea alfabético, numérico o alfanumérico.

Ejemplos de fuentes organizadas según los cuadros de clasificación:

- Oko http://oko.zrc-sazu.si/
- RECERCAT http://www.recercat.net/browse?type=subject
- Open Directory project Dmoz http://www.dmoz.es/
- ISBN http://www.mcu.es/libro/ce/agenciaisbn/infgeneral/tablacdu.html
- The WWW Virtual Library http://vlib.org/
- Librarian's Internet Index http://www.ipl.org/div/subject/index.html
- Buscopio http://www.buscopio.net/esp/
- The www virtual library http://vlib.org/
- Infomine http://infomine.ucr.edu/

Creación de lenguajes nuevos

A partir de un cuadro sistemático podemos elaborar otras clasificaciones o tesauros. Antes de empezar la redacción de un cuadro nuevo, una buena práctica es buscar cadenas ya construidas en otros lenguajes. En la bibliografía científica sobre este tema encontraréis muchos casos resueltos, uno de los cuales es el caso que crea un cuadro de clasificación nuevo para un fondo de economía a partir de la CDU y la JEL http://redc.revistas.csic.es/index.php/redc/article/view/673.

Ya sabemos que en la historia de las clasificaciones documentales los cuadros de clasificación se basan unos en otros, no hay auténticas revoluciones, sino evoluciones. Como dice Jacques Maniez (1992), "en clasificación, como en cualquier disciplina, es inútil reinventar la rueda".

Recuperación con sistemas de clasificación

Los sistemas de clasificación se usan en la Red en catálogos colectivos, bases de datos bibliográficas, directorios de recursos web y portales. Son excelentes para la indización por materias; ahora bien, ¿resultan útiles en la recuperación?

Los últimos años se ha puesto el énfasis en recuperar por palabras clave y conceptos y en sistemas desarrollados, para que sea el usuario quien busque la información sin ayuda profesional, dos características que van en detrimento de los sistemas de clasificación, ya que indizan por materias y requieren ciertos conocimientos técnicos para su utilización.

El uso de la clasificación a la hora de formular búsquedas en los catálogos ha estado habitualmente bastante restringido. En muchos catálogos en línea, la firma decimal solo se utiliza como indicador topográfico, y no se puede indizar o buscar por completo.

La investigación bascula entre mejorar la calidad de los sistemas de clasificación y eliminarlos de la búsqueda.

Opciones de búsqueda

¿Cómo se puede mejorar la recuperación a partir de un cuadro de clasificación? Los autores dan respuestas diversas, algunas de las cuales exponemos a continuación.

1) Convertir el lenguaje codificado y jerárquico en uno natural y alfabético: en otras palabras, convertir las tablas de la CDU en un índice alfabético de términos, de modo que el usuario busque en lenguaje natural *astronomía*, por ejemplo, y el programa responda *52* o dé directamente por pantalla todos los ítems clasificados con 52 y compuestos. Ia McIlwaine define este índice alfabético como un diccionario mediante el cual el usuario puede acceder a la disposición sistemática de las tablas. Construirlo no es una tarea mecánica, exige la aportación intelectual del indizador, porque en este índice hay muchos conceptos compuestos, lugares dobles y sinónimos que habría que definir.

Al final de la CDU se encuentra el índice alfabético, en el que se relaciona el concepto con el código (por ejemplo, Granadura 746.5), pero no nos resulta útil porque este índice es completo, mientras que el SID puede haber escogido un grado inferior de precisión, de detalle. Habría que recortar las clases para que fueran un reflejo de las que se usan en el SID y así no dar la falsa impresión de que la colección es más grande y diversificada. También se deberían abrir los conceptos compuestos (ej.: 543.272 Absorción selectiva de gases), introducir dos veces las remisiones (ej.: 331.215 Salario véase 330.59 Nivel de vida) y abrir los lugares dobles (ej.: 520.1 Observatorios véase 727.912 Arquitectura de observatorios). Si el SID tuviera esta “traducción” de las tablas a una lista alfabética, le resultaría muy provechoso. No se recuperaría directamente por medio de un sistema de clasificación, sino a partir de un sistema de clasificación.

Observación

No debemos confundir este índice de materia con el fichero de autoridades, que es el fichero en que se registran todas las prácticas y decisiones clasificatorias locales.

2) Navegar arriba y abajo del cuadro jerárquico: el usuario visualizaría el árbol de conceptos y escogería el tema deseado. Presenta la ventaja de situar el término en contexto. Por ejemplo, el observatorio de la clase 52 hace referencia a la astronomía; en cambio, el observatorio de la clase 72 hace referencia a arquitectura. Hay webs que deciden no poner los códigos de la notación en la arborescencia para resultar más amigables.

3) Crear redes semánticas a partir de las tablas: una red semántica es conceptualmente muy parecida a un tesauro, muestra los términos en el contexto de sus relaciones semánticas. Ofrece diferentes capacidades de navegación por medio de dispositivos gráficos que representan espacios multidimensionales, referencias cruzadas y notas de alcance.

4) Flexibilizar las búsquedas abriendo las notaciones en partes correspondientes a clases y a facetas (tiempo, lugar, materia, etc.). Los usuarios ya no dependerían de un orden de citación lineal y podrían buscar por partes; por ejemplo, "todo documento que contenga la faceta Francia". Las búsquedas no se tendrían que llevar a cabo con números, sino con términos en lenguaje controlado que remitirían a números que sí reconocería el ordenador.

Ventajas e inconvenientes de recuperar con sistemas de clasificación

Argumentos a favor y en contra de la recuperación con sistemas de clasificación.

Argumentos a favor	Argumentos en contra
• Facilitan la navegación a usuarios sin experiencia. • Proporcionan búsquedas genéricas y específicas. • Permiten localizar los términos de búsqueda en su contexto. • Hacen una partición lógica de largas listas sistemáticas en partes más asumibles. • Dan acceso multilingüe a la colección. • Permiten un uso compartido que mejora la navegación entre varias bases de datos. • Ofrecen estabilidad. • Proporcionan familiaridad. • Dan buen resultado con los números cuando los combinamos con lenguaje natural, porque eliminan ambigüedades, falsas asociaciones y ruidos. • Agrupan los documentos de la misma disciplina, en lugar de quedar dispersos como sucede en la búsqueda alfabética.	• El funcionamiento de la sintaxis del lenguaje se tiene que conocer. • El usuario debe saber la notación del tema. • La actualización es lenta. • El mantenimiento es caro. • La implementación es cara y lenta. • La notación no siempre expresa jerarquía.

Clasificación y categorización

La **Wikipedia**, nacida en el 2001, es una enciclopedia libre mantenida por la Fundación Wikimedia, una organización sin ánimo de lucro. Sus más de quince millones de artículos han sido escritos de forma colaborativa por usuarios de todo el mundo.

Durante los primeros años, la recuperación de información se basó únicamente en los motores de búsqueda y en seguir los enlaces simples entre los artículos. En el año 2004, Wikipedia introdujo el concepto de categorías: cada autor tenía que asignar una categoría a su artículo. El cambio es sustancial por lo que respecta a la indización de esta fuente, ya que en principio funcionaba como una indización con **descriptores libres** (una folksonomía o indización social) y, al evolucionar, optó por combinarlo con una **taxonomía**.

> Esta taxonomía se diferenciaba de las tradicionales en que no estaba hecha a priori y por expertos, sino que, al igual que la indización social, la iban creando los propios autores.

¿En qué consiste la categorización? Se trata de una herramienta del programa MediaWiki que permite almacenar artículos y otras páginas en categorías.

Web recomendada

Los autores Akdag Salah, Gao, Suchecki y Scharnhorst (2010) comparan la CDU con el sistema de categorías de la Wikipedia en el artículo "The Need to Categorize: A Comparative Look at Categorization in Wikipedia and the Universal Decimal Classification System" (http://hth.eccs2010.eu/abstracts.htm#Akdag-Salah-te-al).

Las categorías tienen subcategorías (más específicas) y supercategorías (más generales), que permiten navegar desde lo más general a lo más concreto y a la inversa a partir de una estructura en árbol.

Todos los artículos tienen que pertenecer como mínimo a una categoría. Podemos saber la categoría de cada artículo yendo a la parte inferior de cada entrada de la enciclopedia:

Al final del artículo *Escolástica* encontramos que pertenece a las categorías Filosofía de la edad media y a Teología.



Webs recomendadas

Para más información sobre las categorías en la Wikipedia, leed las páginas de categorización:

http://ca.wikipedia.org/wiki/viquip%C3%A8día:Categoritzaci%C3%B3

Y sobre la sobrecategorización, consultad:

http://es.wikipedia.org/wiki/wikipedia:Sobrecategorizaci%C3%B3n

Para ver todas las categorías, es decir, ir a la raíz de la clasificación de la Wikipedia, hay que hacer clic sobre el logo y acceder a la página principal.





Desde cualquier página de la Wikipedia, simplemente haciendo clic sobre el logo, se accede a la página principal con el índice de categorías.

La clasificación realizada por los wikipedistas se diferencia del etiquetado con *tags* en el hecho de que hay cierto control sobre las **relaciones de significado**. Las páginas de ayuda y los manuales de usuario dan instrucciones a los autores sobre cómo asignar las categorías y cómo evitar la categorización redundante.

Instrucciones básicas:

- Las categorías tienen que ser esenciales y delimitantes, no se pueden crear categorías accesorias o subjetivas.
- Antes de crear una categoría nueva, hay que comprobar si ya existe y si consta con un sinónimo o un nombre similar.
- Se tiene que procurar evitar la sobrecategorización[9] o categorización redundante: no hay que que colocar un artículo en dos categorías cuando una ya contiene la otra.

(9)El autor del artículo *Hepatitis A* lo agrega a la categoría *Enfermedades hepáticas y biliares*, pero no hace falta que lo agregue a la categoría *Hígado* porque ya se incluye en *Enfermedades hepáticas y biliares*.

Reflexión

Fijaos en que, en el caso de la Wikipedia, la clasificación no está hecha ni por un profesional ni por un programa, sino por un tercero, el autor del artículo de la enciclopedia. Veremos este mismo caso en la indización social.

2.3.2. Indexar y recuperar con listas de encabezamientos y listas de autoridades

Las listas de encabezamientos de materia son el segundo lenguaje documental precoordinado que estudiamos, después de los sistemas de clasificación. Igual que estos, disponen de un vocabulario controlado y una sintaxis que precoordina los términos en el momento de la indización.

Las listas de **encabezamientos de materia** son lenguajes naturales, controlados, precoordinados, alfabéticos y que indizan por materias.

Las listas de **autoridades** son lenguajes naturales, controlados, poscoordinados, alfabéticos y que indizan por conceptos.

Listas de encabezamientos de materia y de autoridades en la Web

En la Web disponemos de numerosas listas de encabezamientos de materia y de autoridad, la mayoría mantenidas por bibliotecas nacionales. A continuación, encontraréis las principales listas clasificadas por su idioma.

Observación

El término *encabezamiento* es una traducción literal del inglés *subject headings*. En francés, *vedette-matière*.

1) En catalán:

- Lista de encabezamientos de materia en catalán (LEMAC) (www.bnc.es/lemac/)
- LENOTI (www.bnc.es/lenoti/)
- Biblioteca de Catalunya. Lista de encabezamientos de materia en catalán [en línea]. [Fecha de consulta: 1 de septiembre del 2009.]

Encabezamientos de materia

Las listas de encabezamientos de materia son lenguajes naturales, controlados, precoordinados, alfabéticos y que indizan por materias.

2) En español:

- AM BNE (http://catalogo.bne.es/uhtbin/authoritybrowse.cgi)
- CSIC Autoridades de materia (http://aleph.csic.es/f?func=hilo&hilo_name=find-b&local_base=MAD10)
- Lista de encabezamientos para las bibliotecas públicas (www.mcu.es/bibliotecas/mc/lembp/index.html)

3) En inglés:

- Library of Congress Subject headings 1909 - hasta la actualidad (http://authorities.loc.gov/)
- Bilindex (www.bilindex.com/)

4) En francés:

- Laval Répertoire de vedettes-matière (RVM) de la Universidad Laval (www.bibl.ulaval.ca/mieux/chercher/ch_vedettes_matiere)
- RAMEAU (http://rameau.bnf.fr/)

Elementos de una lista

Los elementos que forman parte de una lista son los encabezamientos y subencabezamientos de materia, las autoridades, las relaciones semánticas, los tipos y la sintaxis. Pasamos a describirlos de forma detallada.

Encabezamientos y subencabezamientos de materia

Una lista de encabezamientos está formada por encabezamientos y subencabezamientos, que pueden ser **simples** o **compuestos**.

Observación

Los términos *encabezamiento* y *epígrafe* son sinónimos. El primero es la traducción literal de *headings*, en inglés, y el segundo es el término en español que propusieron Jorge Aguayo y Carmen Rovira.

Los términos *subencabezado* y *subdivisión* también son sinónimos.

Ejemplos de encabezamientos y subencabezamientos simples y compuestos.

	Encabezamiento	Subencabezamiento
Simple	Alpes	Lesiones
Compuesto	Alpes Dolomitas	Accidentes y lesiones

Algunos subencabezamientos solo se pueden combinar con un encabezamiento concreto, y en estos casos se desarrollan las combinaciones en el mismo epígrafe. En otros casos (la mayoría), los subencabezamientos se pueden com-

binar con un segmento de encabezamientos que cumplan una condición concreta, y así nos lo indica la lista: bajo nombres de persona, bajo guerras, bajo temas, etc.

Ejemplo de subencabezamientos solos o combinados

Encabezamiento + subencabezamiento	Subencabezamiento solo
Alpinismo-Accidentes y lesiones (*Subd. geog.*) No encontraremos Accidentes y lesiones como subencabezamiento solo que podamos combinar con otros encabezamientos.	**Despido** (*Subd. geog.*) Nota de alcance: bajo grupos de profesionales y tipos de empleados.

Autoridades

La materia es una autoridad, pero hay más: nombres propios, congresos, títulos, nombres propios y títulos, entidades y nombres geográficos.

Listas de autoridades

Las listas de autoridades son lenguajes naturales, controlados, poscoordinados, alfabéticos y que indizan por conceptos.

Las autoridades se pueden combinar con todos los lenguajes documentales poscoordinados.

Habitualmente, las listas de encabezamientos de materia solo recogen materias y mantienen el resto de las autoridades en otro fichero y categorizadas en nombres personales, congresos, geográficos u otras etiquetas para identificar el contenido.

Así, si buscamos un nombre propio en la LEMAC, tenemos que hacer clic sobre LENOTI, y si lo buscamos en la BNE o en el CSIC, desplegaremos las opciones para elegir persona y autor personal, respectivamente.

Estas autoridades sirven para representar el nombre del autor o el título uniforme de una obra, pero para nosotros, que estudiamos análisis de contenido, nos son muy útiles, porque la materia de un documento también puede ser:

- **un nombre propio:** documento sobre la vida de William Shakespeare;
- **un nombre de institución presente o histórica:** documento sobre el congreso de Viena 1814-1815;
- **un título uniforme:** interpretaciones de la obra *Fortunata y Jacinta*;
- **un lugar geográfico:** documento sobre Holanda;
- **un nombre de empresa**, **entidad**, **etc.:** documento sobre la fábrica AEG.

Estos términos (*Shakespeare*, *Congreso de Viena*, *Fortunata y Jacinta*, *Holanda* y *AEG*) no aparecen en las listas de encabezamientos de materia porque tienen entrada como nombres personales, de título, geográficos, etc.

Registro de autoridad

La descripción de cada autoridad, con los términos descartados, las referencias, las notas de aplicación y la fuente, se conoce como registro de autoridad. La suma de todos los registros se denomina fichero de autoridades o lista de autoridades. Si este fichero se encuentra vinculado al catálogo bibliográfico, se conoce como catálogo de autoridades. Las autoridades también son conocidas como encabezamientos e identificadores.

Relaciones semánticas

Las relaciones semánticas son las de equivalencia, jerarquía y asociación.

Encabezamiento de *Teatro* de la lista de la BNE

Equivalencia: de un término sinónimo al término aceptado		**Usado por:** Representaciones teatrales Teatro – Representaciones
Jerarquía	Genérica	**Término genérico:** Espectáculos
	Específica	**Término específico:** Adaptaciones teatrales Ballet Mimo Pantomima Sombras chinescas Teatro alternativo Teatro de calle Teatro de marionetas Teatro de variedades
Asociación: evoca otros encabezamientos que podrían ser útiles en la búsqueda		**Término relacionado:** Actores Arte dramático Compañías teatrales Crítica teatral Directores de teatro Escuelas de arte dramático Festivales teatrales Industria del espectáculo Teatro (Género literario) Teatro y sociedad Teatros

Los encabezamientos pueden llevar notas de aplicación que ayudan a definir y matizar el significado. Como indica Martínez Tamayo (2009), las NA (notas de aplicación) pueden ser de cuatro tipos:

1) De definición del epígrafe

Teatro Género literario

Teatros Instalaciones destinadas a la representación teatral

2) Explicativas del alcance del epígrafe

Ecumenismo

Bajo este epígrafe se encuentran las obras sobre la unión de todas las confesiones cristianas [...].

3) Explicativas sobre el uso del epígrafe

Flores

Puede subdividirse geográficamente

4) Nota histórica

Burkina Faso

Epígrafe creado en 1984. Sustituye al epígrafe Alto Volta.

Tipos (según el alcance temático)

Las listas pueden ser de dos tipos:

- enciclopédicas (o universales o generales) y
- especializadas.

Las primeras comprenden todos los ámbitos de conocimiento con una descripción más sucinta, mientras que las segundas tratan con más detalle y más relaciones un tema concreto. Aun así, hay que decir que las listas enciclopédicas tienen una mayor difusión que las especializadas, ya que para indizar de manera específica se prefiere los tesauros.

Observación

Todas las listas del principio de este apartado son enciclopédicas. MESH sería una lista especializada (en medicina y ciencias bioquímicas).

Sintaxis: la precoordinación

Para precoordinar los encabezamientos y subencabezamientos, la regla acostumbra a ser (extracto de Martínez Tamayo, 2009) la siguiente, salvo que la lista indique lo contrario:

Ejemplo de sintaxis de un encabezamiento de materia compuesto

Encabezamiento	Subdivisiones			De forma o género
	De tema	Geográficas	Cronológicas	
Alpes	Clima	Argentina	1952	Informe

Encabezamiento construido: Alpes – Clima – Argentina – 1952 – Informe

Recuperación con listas de encabezamientos y autoridades

La recuperación con estos lenguajes es muy asequible: los catálogos nos proporcionan una lista previa de términos de indización para escoger a la carta, la red de relaciones semánticas nos ayuda mucho a encontrar otros documentos a partir del tema que nos interesa y son lenguajes totalmente automatizados,

hasta el punto de que un encabezamiento compuesto se puede recuperar a partir de las piezas que lo forman (encabezamientos y subencabezamientos) como si fueran palabras clave. Y una cuarta característica es que son los mejores traductores para buscar listas de otros idiomas.

Resultado directo o lista de encabezamientos para escoger

El proceso de búsqueda es sencillo: escribimos el tema que buscamos y el catálogo nos devuelve dos opciones. Fijémonos en que la primera devuelve documentos y la segunda, una pantalla intermedia de encabezamientos para refinar la búsqueda:

1) Los documentos indizados con el término que hemos escrito. Esto sucede cuando la demanda coincide exactamente con el documento indizado.

Biblioteca de la UOC: el usuario busca *arte prehistórico* y recupera cuatro documentos directos:

La Imagen de la mujer en el arte prehistórico
Delporte, Henri
Madrid: Colegio Universitario. Ediciones Istmo, DL 1982
La Imagen de los animales en el arte prehistórico
Delporte, Henri
Madrid: Compañía Literaria, 1995
Manual de arte prehistórico
Sanchidrián Torti, José Luis
Barcelona: Ariel, 2001
Les Origines de l'art
Lorblanchet, Michel
París: Pommier: Cité des sciences et de l'industrie, cop. 2006

Esto quiere decir que, a pesar de que el encabezamiento *Arte prehistórico* se puede abrir en otros encabezamientos (por ejemplo, con subdivisiones geográficas tipo *Arte prehistórico – Francia*), la biblioteca no dispone de más documentos.

2) Una lista de los encabezamientos y subencabezamientos compuestos a partir de nuestra petición. Esto sucede porque hay documentos indizados de manera compuesta a partir del término que hemos pedido y el catálogo nos ofrece la posibilidad de refinar la búsqueda.

Pongamos un ejemplo de la Biblioteca de la UOC: el usuario busca por archivos, pero hay tantos documentos compuestos que nos devuelve los encabezamientos con el número de documentos de cada uno:

Archivos	15
Archivos - Actividades culturales	1
Archivos - Administración	4
Etc.	

Observación

De este ejemplo se puede deducir que la precoordinación ofrece una mayor carga informativa que la poscoordinación, especialmente en los encabezamientos compuestos, es decir, a fuerza de ir uniendo subdivisiones, el contenido se matiza hasta ofrecer una idea más precisa del documento.

Referencias de las listas y su utilidad en la recuperación

Los tres tipos de relaciones semánticas –equivalencia, jerarquía y asociación– ayudan a ampliar las posibilidades de encontrar más documentos.

1) Las relaciones de equivalencia son básicas, ya que es posible que se haga la búsqueda por un término no aceptado (por sinónimos, siglas, términos demasiado especializados, barbarismos, formas incorrectas, etc.).

Estas relaciones son invisibles para el usuario, incluso es posible que no se dé cuenta de que ha buscado por el término *localizaciones cinematográficas* y el catálogo le haya respondido con documentos sobre Cinematografía – Exteriores. Es un proceso automatizado.

2) Las relaciones de jerarquía pueden ser de dos tipos: de términos genéricos o de términos específicos. El papel de cada una en la recuperación es diferente, ya que mientras que una crea ruido, la otra es efectiva:

a) Las relaciones del término que se está buscando frente a un término genérico: crean ruido, ya que amplían el campo de alcance del concepto.

El usuario busca por *gospel*, no encuentra documentos interesantes y decide buscar por un término genérico como *cantos sacros*. El resultado serán documentos que tratarán parcialmente de su tema de interés original.

b) Las relaciones del término que se está buscando frente a un término específico: son correctas. Amplían la recuperación a documentos que entran de lleno en el tema que se está buscando y ofrecen mucha información detallada (a menudo incluso demasiada).

Buscamos por *música country* y recuperamos documentos sobre *bluegrass*, *música country rock*, *rockabilly* y *western swing*.

3) Las relaciones de asociación también son muy importantes, ya que nos amplían de una manera diferente el abanico de posibilidades: si las relaciones jerárquicas anteriores sitúan el término en una posición vertical (mayor o menor que), ahora las relaciones asociativas lo relacionan de manera horizontal. Los términos asociados se evocan el uno al otro, están relacionados y mentalmente los conectamos. La relación entre ellos es simétrica y, por consiguiente, recíproca.

Buscamos por *justicia* y la lista nos sugiere buscar por *derecho natural*, *igualdad ante la ley*, *justicia distributiva*, *justicia social* y*justicia transicional*.

Descomposición de los encabezamientos en palabras clave de materia

En los sistemas de clasificación se apuntaba que una de las maneras de mejorar la recuperación sería particionar la notación en facetas, en partes autónomas que fueran buscables. Lo que en aquel lenguaje era un tema pendiente, en las listas de encabezamientos es un asunto resuelto.

A la hora de indizar, precoordinamos los términos en un encabezamiento compuesto, pero a la hora de la recuperación podemos buscar las partes por separado gracias a la opción de indizar la palabra clave de materia.

Haciendo un juego de palabras, primero controlamos y después descontrolamos. En la LEMAC, hay que pedir el campo palabra clave de materia; en la lista del CSIC, ya viene por defecto.

Ejemplo

Primero controlamos un encabezamiento como *Universidades – Archivos*, porque *archivos* es un subencabezamiento que, tal como indica la lista, se puede usar detrás de entidades, pero a la hora de buscar podemos pedir solo *archivos* y el catálogo nos devolverá también los documentos indizados como *Universidades – Archivos*.

Otra utilidad de buscar por palabra clave de materia es que podremos recuperar el término que buscamos que se encuentre en la posición del encabezamiento.

Si buscamos por *archivos*, el programa buscará por la *A*, pero no recuperará documentos indizados con un encabezamiento como *Documentos de archivos*. En cambio, si se busca *archivos* como palabra clave, se recuperan todos los encabezamientos.

Finalmente, otra opción muy útil que ofrece la lista de autoridades del CSIC es que hace la búsqueda de la palabra clave en cualquier posición, ya no dentro del encabezamiento como en el caso anterior, sino dentro del registro de cada autoridad.

En la búsqueda de *archivos* en el CSIC por materia, en el resultado podemos observar que hay encabezamientos en los que la palabra *archivos* no aparece y, en cambio, la tenemos en pantalla. En la columna de al lado explicamos el motivo para ello.

Universidades-- Archivos	Aparece en el encabezamiento
Registros eclesiásticos	Es un término asociado
Prueba (Derecho)	Es un término asociado
Protocolos notariales	Es un término asociado
Manuscritos	Es un término asociado
Informática-- Documentación	Es un término equivalente
Fototecas	Es un término equivalente
Documentos de archivos	Aparece en el mismo encabezamiento
Documentos administrativos	Es un término asociado
Diplomática	Es un término asociado

Si ampliamos el registro de *Diplomática*, comprobamos que efectivamente aparece *archivos* (es un término asociado).

N sistema	000028986
Encabezamiento	● Diplomática
Térm. genérico	Cartularios
Térm. genérico	Historia
Térm. genérico	Historiografía
Térm. específico	Sigilografía
Térm. específico	Notas tironianas
Véase además	Archivos
Véase además	Manuscritos
Véase además	Paleografía
Fuente	LCSH (Diplomatics)
En LCSH/BNF	● Diplomatics

Encabezamientos en otros idiomas

Una de las informaciones que consta en la ficha de cada autoridad de materia es la fuente original de la que se ha importado el concepto. La mayoría de las listas de materia se basan unas en otras; las inglesas LCSH y la francesa RAMEAU son las más influyentes en el ámbito internacional.

Cuando se tienen que efectuar búsquedas en otros idiomas, se recomienda consultar primero una lista en el idioma propio y comprobar el nombre original. Es el mejor traductor que hay.

El concepto *servicios de resúmenes*, que resulta ser un término no aceptado, es *Abstracting and indexing services* en inglés y *services d'analyse et d'indexation des documents* en francés.

N.° sistema	0000045211
Encabezamiento	● Servicios de análisis documental
Usado por	● Análisis documental, Servicios de
	● Centros de resúmenes
	● Servicios de *abstracts*
	● Servicios de extractos
	● Servicios de índices
	● Servicios de indización
	● Servicios de resúmenes e indización
Térm. genérico	Bibliografía
Térm. genérico	Documentación
Térm. genérico	Índices
Térm. específico	Sistemas de información
Fuente	LCSH (Abstracting and indexing services)
	BNF aut. en línea (21/02/07) (Services d'analyse et d'indexation des documents)
	BNE aut. en línea (21/02/07) (Servicio de resúmenes)
En LCSH/BNF	Abstracting and indexing services
	● Services d'analyse et d'indexation des documents

2.3.3. Indexación y recuperación con tesauros

Indizar con un tesauro, igual que con todos los lenguaje documentales poscoordinados, es muy sencillo. Se trata de lenguajes en los que no hay sintaxis; por lo tanto, la dificultad no estriba en la composición, el orden y la sintaxis del término de indización, sino en la selección de los descriptores.

Tesauros

Los tesauros son lenguajes naturales, controlados, poscoordinados, jerárquicos y alfabéticos y que indizan por conceptos.

Tesauros en la Web

Existe un gran número de tesauros en línea y gratuitos en la Red. Encontramos tesauros de agricultura, astronomía, biblioteconomía, biología, arte, etc. A continuación, os ofrecemos una selección clasificada por temas.

Lista de tesauros en línea.

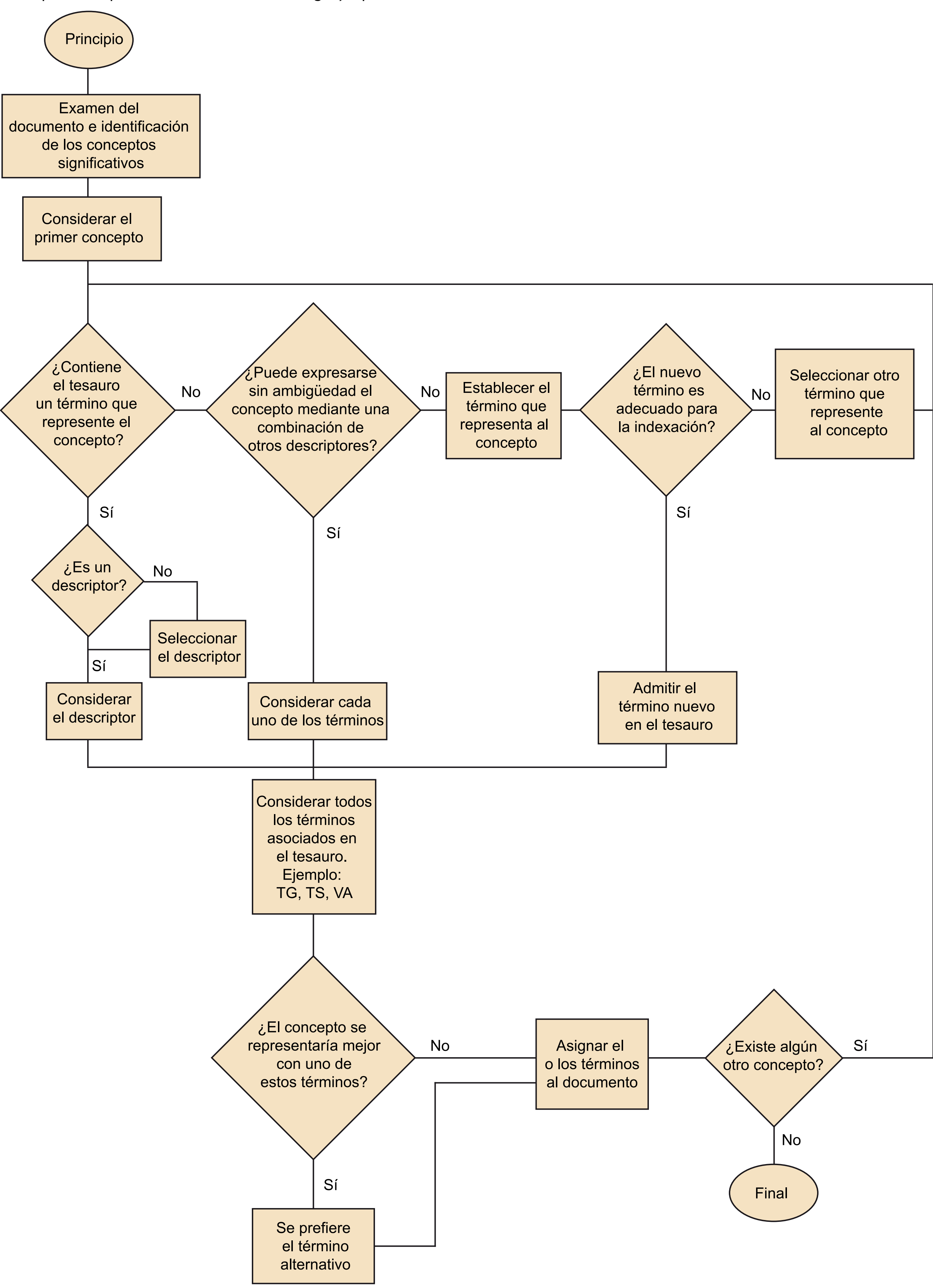
Temática	Nombre del tesauro
Agricultura	AGROVOC
Astronomía	The Astronomy Thesaurus
Biblioteconomía	IEDCYT - Tesauro de biblioteconomía y documentación DOCUTES Universidad de León
Biología	IEDCYT - Tesauro de biología animal
Ciencia	IEDCYT - Tesauro SNIPES
Demografía	Population Multilingual Thesaurus
Economía	EUROVOC Thesaurus IEDCYT - Tesauro ISOC de economía
Educación	EUROVOC Thesaurus
Empresa	EUROVOC Thesaurus IEDCYT - Tesauro de propiedad industrial
Geografía	EUROVOC Thesaurus Getty Thesaurus of Geographic Names IEDCYT - Tesauro de topónimos
Geología	IEDCYT - Tesauro de geología
Historia	IEDCYT - Tesauro de historia contemporánea de España Historia de Catalunya
Lengua y literatura	Traces. Base de datos de lengua y literatura catalanas - Tesauro
Matemáticas	BUCM Tesamat Biblioteca Complutense
Propiedad industrial	CSIC - Tesauro de propiedad industrial
Psicología	IEDYCT - Tesauro ISOC de psicología
Sociología	EUROVOC Thesaurus IEDCYT - Tesauro de sociología
Topónimos	CSIC - Tesauro de topónimos
Urbanismo	IEDCYT - Tesauro de urbanismo
Genéricos	UNESCO Historia de Catalunya Microtesauros temáticos de la UB SPINES del IEDCYT ERIC

La mayoría de los tesauros son especializados, pero algunos son genéricos, como el EUROVOC o los últimos de la lista.

¿Cómo se indiza con un tesauro?

El proceso para indizar con tesauros y, por extensión, con cualquier lenguaje documental poscoordinado se encuentra explicado de forma gráfica en la norma UNE-50-121-91, anexo A, página 7.

Descripción del proceso de indización con lenguajes poscoordinados



+++El indizador examinará el documento y extraerá conceptos de él, conceptos que después traducirá a descriptores del tesauro. Primero se busca en la presentación alfabética y después se comprueba en la presentación jerárquica (esta segunda consulta ayuda a visualizar la posición del descriptor en todo el árbol). Los descriptores que le interesan pueden estar en varias microdisciplinas y en diferentes niveles de sangría.

Ejemplo de descriptores en diferentes microdisciplinas

Documento	Indización
Keefer, Alice (2007). "Los repositorios digitales universitarios y los autoras" [en línea]. *Anales de Documentación* (núm. 10 págs. 205-214). Disponible en http://revistas.um.es/analesdoc/article/viewfile/1151/1201.	Bibliotecas universitarias Fuentes de información Documentos electrónicos Universidades Documentación Bases de datos

Hemos indizado con el Tesauro de Historia de Catalunya (http://sdhlc.uab.cat/tesaurus.htm). Los tres primeros descriptores son de la microdisciplina [Documentación e información]. El que indica Universidades es de [Educación]. Los dos últimos son de [Ciencia y Tecnología].

Ejemplo de descriptores en diferentes niveles de sangría

Documento	Indización
Programa electoral presentado por Convergència i Unió de Sant Andreu de Llavaneres en las elecciones municipales del 2007 y que también contiene la lista de candidatos de este partido.	Partidos políticos Programa electoral Elecciones municipales 2007 Candidaturas electorales Convergència i Unió (proveniente de LENOTI) Sant Andreu de Llavaneres (proveniente de la GEC)

En esta ocasión solo hemos necesitado una microdisciplina, la de política, porque el documento no hace referencia a otros temas.

Nombres propios y geográficos

Recordemos que los nombres propios y el geográficos no se encuentran en el tesauro, sino que provienen de listas de autoridades como las del ejemplo (LENOTI y Gran Enciclopèdia Catalana).

1 [POLÍTICA]

[Acción política]
. Vida política
.. Oposición política
... Atentados
... Bullangas NA: [1835-1842]
... Conspiraciones NA: utilizarlo acompañado de la fecha de la conspiración.
... Guerrillas
.... Guerrilleros
.... Maquis
... Insurrección
.... Insurrección federal NA: utilizarlo acompañado de la fecha.
... Oposición parlamentaria
... Rebeliones militares
.... Golpe de estado
.... Pronunciamiento
... Terrorismo

.. Partidos y grupos políticos NA: no utilizarlo como descriptor.
... Asociaciones políticas
... Partidos políticos

.. Represión política
... Depuraciones políticas
... Exilio
... Persecuciones políticas
... Prisioneros políticos

.. Sistema electoral
... Elecciones
.... Abstencionismo electoral
.... Campañas electorales
.... Candidaturas electorales
.... Comportamiento electoral
.... Elecciones autonómicas NA: utilizarlo acompañado de la fecha de celebración.
.... Elecciones generales NA: utilizarlo acompañado de la fecha de celebración.
.... Elecciones municipales NA: utilizarlo acompañado de la fecha de celebración.
.... Elecciones Parlamento Europeo NA: utilizarlo acompañado de la fecha de celebración.
.... Elecciones provinciales NA: utilizarlo acompañado de la fecha de celebración.
.... Fraude electoral
.... Programa electoral
.... Referéndum NA: utilizarlo acompañado de la fecha de celebración.
... Insaculación
... Sufragio universal
2

En primer lugar, conviene fijarse en que los descriptores seleccionados forman parte de cadenas jerárquicas diferentes. Un error sería indizar *Elecciones* porque es el término amplio (TA) de *Candidaturas electorales*, *Elecciones municipales* y *Programa electoral*. No podemos indizar el descriptor (o término) específico (TE) y su TA al mismo tiempo.

Reflexión

Esta es la única regla que necesitamos conocer para indizar con tesauro: no indizar el TA y el TE a la vez.

En segundo lugar, conviene prestar atención al hecho de que hay que ajustar el enunciado al descriptor aprobado y admitido en el tesauro: *lista de candidatos* por *Candidaturas electorales*.

En el proceso de mantenimiento de un tesauro, es posible que conceptos no recogidos en un primer momento se acaben añadiendo posteriormente, pero esta tarea corresponde al administrador del tesauro y no al documentalista; en todo caso, el documentalista puede proponer la necesidad de un descriptor nuevo en un campo denominado *Descriptores candidatos*.

Creación de un tesauro

Los tesauros tienen las presentaciones básicas de todo lenguaje documental: la jerárquica, la alfabética, la gráfica y la permutada.

Recordemos que las fases de construcción de un tesauro son ocho en los monolingües y nueve en los multilingües.

1) Recogida del vocabulario en lenguaje natural dentro del dominio que incluirá el tesauro.

2) Subdivisión del conjunto de los dominios que se tendrán en cuenta en una serie de microdisciplinas.

3) Transformación del vocabulario libre en un lenguaje controlado, establecimiento de las relaciones de pertenencia, equivalencia semántica y jerarquía y redacción de las notas explicativas.

4) Búsqueda de las equivalencias interlingüísticas (si se trata de un tesauro multilingüe).

5) Enriquecimiento del tesauro por medio de relaciones asociativas.

6) Elaboración del borrador del tesauro.

7) Formación de los indicadores.

8) Prueba del tesauro.

9) Revisión final y primera edición.

Los descriptores de cada microdisciplina pueden estar ordenados de tres maneras diferentes:

- cronológicamente,
- alfabéticamente, o
- según el proceso.

Los dos primeros criterios son claros, el tercero se refiere a procesos que ya tienen un orden lógico interno como, en el ejemplo, el orden de los estudios: primero preescolar, seguido de primaria, secundaria y superior.

Reflexión

Si sabemos construir un tesauro, sabemos construir todos los lenguajes documentales. Además, al ser especializado, es el lenguaje perfecto para construirlo a la medida de nuestras necesidades. Por todos estos motivos, pues, resulta conveniente saber construir un tesauro.

Lecturas recomendadas

Para más información sobre el proceso y las fases, recomendamos las lecturas siguientes: Aitchison (1987), Lancaster (2002), Slype (1991) y las normas UNE 50-106 (ISO 2788-1986) y UNE-50-125 (ISO 5964-1985).

Tres tipos de ordenaciones

Cronológicamente	Alfabéticamente	Según el proceso
[Historia contemporánea] .Edad contemporánea .. Crisis del Antiguo Régimen NA: [1808-1833] … Guerra de la Independencia Española NA: [1808-1814] ….Batallas de El Bruc NA: [1808]	. Medio ambiente .. Residuos ... Aguas residuales ... Residuos industriales	[Niveles de enseñanza] .. Preescolar .. Educación primaria .. Educación secundaria .. Educación superior

Finalmente, apuntamos que las facetas de un tesauro se pueden ordenar según la conveniencia de los constructores para que resulten más claras, como por ejemplo las facetas de la microdisicplina de [ECONOMÍA], en las que vemos que *Economía general* precede al resto.

- [Historia económica]
- [Economía general]
- [Economía agraria]
- [Economía pesquera]
- [Economía industrial]
- [Comercio]
- [Hoteleria y turismo]
- [Finanzas]
- [Economía de la empresa]

Recuperación con tesauro

La recuperación con un lenguaje analítico y poscoordinado como los tesauros es más sencilla que la de lenguajes precoordinados porque no hay sintaxis y se pueden añadir tantos descriptores como se considere oportuno.

Igual que en la indización, es muy importante que el indizador conozca de forma exhaustiva el tesauro que indiza la base de datos, las microdisciplinas y el alcance conceptual de cada una. También es preciso que conozca las listas de autoridades de su SID, tanto por nombres geográficos como personales, títulos o entidades.

Proceso de búsqueda con un tesauro

El proceso de búsqueda con tesauro consta de tres partes:

- recogida de conceptos,

- traducción al lenguaje, y
- formulación de la búsqueda.

Ejemplificaremos una búsqueda en la base de datos ISOC – Biblioteconomía y documentación, a partir del tesauro de Biblioteconomía de la IEDCYT (IEDCYT – Tesauro de Biblioteconomía y Documentación).

Recogida de conceptos

El tesauro es un lenguaje documental analítico y, como tal, permite pedir tantos descriptores como sea necesario. Es importante que la petición de información se formule de manera exhaustiva con el fin de recoger todos los conceptos interesantes para el usuario y que podemos encontrar idénticos o no en el tesauro.

El usuario pide documentación sobre documentos de archivo de oficina a la empresa y el documentalista acota la petición a los descriptores que conoce de su tesauro.

¿Qué tipo de empresa, pública o privada? ¿De qué sector? ¿Documentos contables? ¿Normativas? ¿Cómo clasificarlos? ¿Política de expurgo? ¿De qué años? ¿Todo tipo de documentales? ¿Todos o solo un segmento? Etc.

Traducción al lenguaje

Una vez que el documentalista dispone de los conceptos, la segunda tarea es localizarlos en el tesauro para traducirlos. Aquí el documentalista jugará con las tres presentaciones básicas de todo tesauros: la alfabética, la jerárquica y la permutada.

El documentalista se puede encontrar en dos situaciones: encuentra el concepto expresado más o menos de la manera que pensaba o bien no lo encuentra.

1) Para localizar el descriptor, hay que consultar la **presentación alfabética** del tesauro. En un primer momento se consulta esta presentación y no la jerárquica por los motivos siguientes:

a) Porque la presentación alfabética tiene las relaciones de equivalencia entre el no-descriptor y el descriptor aceptado.

En la expresión del usuario era *Archivos de oficina*, que es un no-descriptor que remite a *Archivos de gestión*.

b) Para comprobar cómo se escribe el descriptor, es decir, cuál es la forma aceptada.

En la expresión del usuario era *Archivos de oficina en la empresa* y en el tesauro el concepto se formaliza en *Archivos de empresas*; *Archivos de gestión*.

c) Porque el documentalista no sabe a qué microdisciplina o faceta pertenece el descriptor.

> *Archivos de empresas* y *Archivos de gestión* no pertenecen a [*Archivística*], sino a la microdisciplina de [*Unidades de información*].

d) Si lo buscara por la sistemática, tendría que repasar el tesauro entero para localizarlo; en cambio, con la alfabética los encontrará a la primera.

Si el documentalista no encuentra el descriptor, entonces le serán más útiles la presentación jerárquica y la permutada.

2) Consultar la **presentación jerárquica**. Su utilidad estriba en el hecho de que la arborescencia le puede sugerir descriptores paralelos, genéricos y específicos. Pondremos un ejemplo de cada uno de ellos.

Ejemplo de términos paralelos

El documentalista busca algún concepto que exprese la cadena documental en *archivos*. No aparece en el tesauro y tampoco es un no-descriptor. Aunque no aparezca, se da cuenta de que todas las fases de la cadena se encuentran sistematizadas bajo el descriptor *Proceso documental*. En una segunda opción, podría abrir el descriptor en términos más específicos y buscar por fases y subfases concretas de la cadena; por ejemplo, *Adquisiciones*; *Análisis de contenido*.

Ejemplo de términos genéricos

El usuario ha preguntado por el concepto *unitérminos*, que no consta en el tesauro y tampoco hay ningún otro término que pueda usar. En este caso, seleccionaría el descriptor inmediatamente superior conceptualmente a otros descriptores paralelos, es decir, si *unitérmino* está en el mismo nivel que *descriptor* y que *palabra clave*, escogería *Términos*, que engloba todos los tipos de términos de indización. Otro caso se da cuando el documentalista encuentra el descriptor correcto; por ejemplo, *Reglamentos de archivos*, pero la base de datos le devuelve cero resultados, por lo que decide consultar la jerárquica y reformular la búsqueda, esta vez con el término genérico de *Reglamentos de archivos*, que es *Política archivística*.

Ejemplo de términos específicos

El usuario ha preguntado por el tema *lenguajes documentales*. El tesauro recoge este concepto como descriptor, pero el documentalista, al consultar la presentación jerárquica, se da cuenta de que también puede buscar por los términos específicos que están en ese tesauro:

TE Clasificaciones

TE Lenguajes de indización

Observación

Recordemos que el documentalista no habrá indizado con el TA y el TE al mismo tiempo. Por lo tanto, un manual general sobre lenguajes documentales estará indizado como *Lenguajes documentales* y no con el descriptor de cada lenguaje concreto.

3) Consultar los **índices permutados**. Los índices permutados (KWIC o KWOC) permiten localizar otros descriptores que contengan la palabra clave que buscamos **en cualquier posición del descriptor**.

Si buscamos *archivos*, además de la letra *A* de *archivos*, si consultamos el índice KWIC podemos encontrar:

Automatización de archivos

Historia de los archivos

Sistemas nacionales de archivos

...

Formulación de la búsqueda

Finalmente, formulará la búsqueda distribuyendo los conceptos en los campos de la base de datos (por materia, alcance cronológico, formato, idioma, etc.) y haciendo uso de operadores booleanos si es preciso.

2.3.4. Indización con listas de descriptores libres: etiquetas e Indización social

La lista de descriptores libres es un lenguaje que se crea dinámicamente, en tiempo real, a medida que el indizador va leyendo y asignando un término. Los términos del vocabulario no constan en ninguna hoja previa; el indizador no comprueba que el término exista ni cómo se escribe. Hay libertad plena.

Listas de descriptores libres

Las listas de descriptores libres son lenguajes naturales, libres, poscoordinados, alfabéticos y analíticos por conceptos.

Descriptores libres en la Web

En la Web existen numerosas iniciativas de indización con descriptores libres; las más meritorias son los marcadores sociales (Delicious), webs para compartir imágenes (Tagzania, Flickr, YouTube) y aplicaciones de la Web 2.0, como blogs (Blogger), redes sociales y webs (Buzzillions), que recogen la opinión de consumidores sobre marcas de todo tipo de productos.

- Delicious (https://www.delicious.com): Diigo (http://www.diigo.com) y Mr Wong (http://www.mister-wong.com) son servicios de gestión de direcciones de interés a través de la Web. Permiten guardar y recuperar en la Red las direcciones de interés, que clásicamente se almacenaban desde el navegador localmente en el ordenador, de forma que son consultables en línea y no solo de forma local.

- Tagzania (http://www.tagzania.com): se trata de un sistema que usa folksonomías sobre la API del potente Google Maps. Es un *mashup* de geolocalización de fotografías similar a Panoramio (http://www.panoramio.com) que ofrece otras funcionalidades de valor añadido a los mapas.

- Flickr (http://www.flickr.com): es un sitio web de Yahoo para organizar fotografías digitales que funciona como una red social. Es un servicio muy utilizado por los usuarios de blogs como depósito de fotos.

- YouTube (http://www.youtube.com): es un sitio web para compartir vídeos, clips de películas, clips de televisión y vídeos musicales, así como

contenido aficionado. Los usuarios no registrados pueden ver vídeos, y los usuarios registrados pueden subir un número ilimitado de vídeos.

- Blogger (https://accounts.google.com): se trata de un servicio para crear y publicar un blog de una forma muy fácil.

- Buzzillions (www.buzzillions.com): es una página web que recoge cerca de diecisiete millones de críticas de productos de una amplia gama de categorías (electrónica, moda, salud, etc.). Las recomendaciones provienen de personas reales (no se pagan por las revisiones), con la intención de asesorar a nuevos compradores a partir del grado de satisfacción de los productos.

Etiquetas e indización social

Cada usuario indiza los descriptores libres que le parecen mejores. Millones de usuarios indizan sus descriptores. Entre todos crean un espacio de aportaciones sin una intervención centralizada ni más autoridad que la que aportan los usuarios, no hay descriptores predeterminados.

James Surowiecki

James Surowiecki (2004) lo denomina la sabiduría de las masas (*the wisdom of crowds*).

Esta forma de indizar, no profesional y sin lenguaje documental controlado, se conoce como **indización social**. En ella intervienen las etiquetas o *tags*, el *tagging* o acción de indizar libremente y las folksonomías o conjunto total de todas las etiquetas asignadas por los usuarios.

Supone una revolución en el mundo de la Web porque se ha invertido el paradigma: antes pocos autores escribían para muchos lectores, y ahora muchos autores no solo escriben, sino que también editan y describen sus documentos.

Como dice Mari Carmen Marcos (2009):

> "cada cual es autor, editor y documentalista a la vez".

Terminología

Encontraremos varios términos para cada concepto:

- Para la indización: *descriptores libres* o *etiquetas* o *tags*. El conjunto de *tags* sedenomina *nube de tags*, que sería lo más parecido a un lenguaje documental.
- Para la acción de indizar libremente: *tagging* o *etiquetado social* y, más específicamente, cuando se trata de describir los recursos web, *social bookmarking* o *website bookmarking*.
- Para el conjunto de *tags* de todos los usuarios: *folksonomías* o *clasificación hecha por el pueblo*.

Etiquetas

Una etiqueta o *tag* es un término de indización que se añade a un objeto digital, como una web, un vídeo o una foto, para describirlo en forma y contenido.

Ejemplo

Por ejemplo, enciclopedia_arte: enciclopedia (forma) de arte (contenido). No es un descriptor controlado, es un descriptor libre.

Las primeras etiquetas aparecieron en los blogs, y proporcionaban enlaces y comentarios sobre recursos del tipo "recomiendo la web tal para tal tema". Se considera que fueron los primeros metadatos, aunque carentes de estructura. Hoy en día, los usuarios indizan con etiquetas sus webs preferidas, las localizaciones de las fotos, las emociones de unas imágenes, el grado de satisfacción de un lavaplatos, etc.

Las etiquetas resultan funcionales porque son las autoridades de los usuarios. Lancaster ya observaba en el año 1995 que los términos se tenían que obtener de los usuarios potenciales y que debían representar sus intereses concretos. O, retrocediendo más en el tiempo, Cutter ya postulaba que los términos de indización tenían que representar el uso común y poner la atención en el lector.

Las etiquetas pueden ser unitérminos o descriptores compuestos, es decir, pueden estar formadas por una sola palabra (tesauro) o por dos palabras (por ejemplo, Lenguajes_documentales).

Separación con guión

Las palabras se acostumbran a separar con guión porque el espacio es el signo que marca el final de la etiqueta.

Ros-Martin (2008) clasificó las etiquetas en los grupos siguientes:

1) Las basadas en el contenido temático.

Ejemplo: Capítulo_indización_social

2) Las basadas en el contexto o almacenamiento.

Ejemplo: Módulo3_cap2

3) Las subjetivas.

Ejemplo: Útil

4) Los atributos que no se deriven del contenido.

Ejemplo: UOC

5) Las de organización o de recordatorio de tareas.

Ejemplo: Guardar, Relacionar_con_Recuperación, Para_Juan

El conjunto de etiquetas se conoce como **nube de etiquetas**. Esta nube es un espacio plano en el que las etiquetas no guardan relaciones de parentesco ni de jerarquía entre ellas pero que permiten la compartición de categorías entre usuarios. Se presentan en orden alfabético y destacadas con una tipografía más grande según la frecuencia de uso.

africa animals architecture art australia autumn baby band barcelona beach berlin bird birthday black blackandwhite blue bw california cameraphone canada canon car cat chicago china christmas church city clouds color concert cute dance day de dog england europe fall family fashion festival film florida flower flowers food football france friends fun garden geotagged germany girl girls graffiti green halloween hawaii hiking holiday home house india ireland island italia italy japan july kids la lake landscape light live london love macro me mexico mountain mountains museum music nature new newyork newyorkcity night nikon nyc ocean old paris park party people photo photography photos portrait red river rock rome san sanfrancisco scotland sea seattle show sky snow spain spring street summer sun sunset taiwan texas thailand tokyo toronto tour travel tree trees trip uk urban usa vacation vancouver

Fuente: imagen tomada de Flickr.

Indización social

Los descriptores libres son el lenguaje ideal para indizar la Web por los factores siguientes:

1) Se trata de un lenguaje libre. La Web no se puede indizar con los lenguajes controlados, porque el tiempo y el esfuerzo económico que se derivarían de ello serían inasumibles. Los lenguajes documentales controlados no son adecuados en entornos en los que los metadatos resultan una opción mejor. Los metadatos pueden ser de varios tipos: creados por un documentalista, por el autor del documento o por un robot. Con las etiquetas podemos añadir otra vía, la de los metadatos creados por los usuarios (Mathes, 2004).

2) No necesitan formación documentalista previa: las características de este lenguaje lo hacen ideal para cualquier colectivo no profesional de la documentación, como es el caso de los internautas.

3) El grupo de usuarios es tan numeroso que es capaz de asumir cantidades ingentes de documentos (ya no hablamos de un indizador, sino de una comunidad de indizadores).

4) Permiten indizar documentos como imágenes o vídeos que no vayan acompañados de texto o de pies de foto, que hasta ahora solo eran indizables por humanos y no por robots.

5) Las etiquetas son cercanas a los usuarios; no son términos escogidos por técnicos, sino que se trata de términos intuitivos. La comunidad actúa como una criba que filtra las palabras realmente más útiles.

6) Son eficaces individualmente –en el ámbito del usuario– porque organizan la información personal y, socialmente, porque toda la comunidad virtual se beneficia de la indización que han hecho los demás.

Lecturas recomendadas

Se han hecho varios estudios sobre la consistencia de indizar con etiquetas entre indizadores a la hora de indizar imágenes e incluso emociones con resultados muy buenos de coherencia entre usuarios (emociones identificadas de manera homogénea). Un ejemplo lo tenéis a Knautz and Stock (2010) y a Ransom and Rafferty (2011):

Kathrin Knautz; Wolfgang G. Stock (2010). "Collective indexing of emotions in videos". *Journal of Documentation* (vol. 67, núm. 6, págs. 975-994).

N. Ransom; P. Rafferty (2011). "Facets of user-assigned tags and their effectiveness in image retrieval". *Journal of Documentation* (vol. 67, núm. 6, págs. 1.038-1.066).

Los profesionales de la información también usan la indización social o *tagging* para indizar los recursos web. Se utilizan en intranets, sistemas corporativos, bases de datos y bibliotecas para dar valor añadido a sus bases de datos (por ejemplo, la base de datos Complured de la Universidad Complutense de Madrid), así como para compartir los marcadores seleccionados con otros usuarios y para reutilizar los contenidos en otras aplicaciones como redes sociales de tipo Twitter, y de este modo proporcionar una mayor visibilidad a la institución.

Organización de las colecciones en las bibliotecas universitarias

La mayoría de las bibliotecas universitarias organiza las colecciones de la manera siguiente:

- **Colección propia:** catálogo indizado de forma controlada (sistemas de clasificación + listas de encabezamientos + lista de autoridades / tesauros + lista de autoridades) y automática (lista de palabras clave).
- **Recursos electrónicos de la Web:** directorios temáticos o guías temáticas (sistemas de clasificación) + Delicious (lista de descriptores libres o*tags*).

Podéis comprobar que las etiquetas de un Delicious son descriptores libres haciendo la siguiente comparación: buscad una lista de encabezamientos de materia que se use o se cree en una biblioteca, y entonces consultad el Delicious de esa biblioteca.

Por ejemplo, la Biblioteca de Catalunya, autora de la LEMAC, indiza en el catálogo con el encabezamiento Arte – Historia, pero Delicious indiza Historia del arte, que es un término más próximo al usuario.

Solo hay que consultar las bibliotecas de universidades que imparten Información y Documentación para darse cuenta de que, además del catálogo, tienen Delicious.

- Delicious de la Universidad de Barcelona, CRAI (www.delicious.com/craiubreferencia).

- Delicious de la Universidad Nacional de Educación a Distancia (UNED) (http://delicious.com/brelreferencia20).

- Delicious de la Universidad Complutense de Madrid (http://delicious.com/bibliotecacps).

Los indizadores tienen varias motivaciones para hacer indización social, ya que obtienen varios beneficios sociales de ello. Javier Cañada (2006) los clasificó tal y como queda recogido en la tabla que tenéis a continuación.

Tipología de motivaciones de las personas a la hora de etiquetar.

Tipo de etiquetado	Beneficio social	Motivación
El etiquetado egoísta: etiquetar en beneficio propio; suelen ser etiquetas muy significativas para el usuario pero no para la comunidad. Ej.: "para_leer".	Si las etiquetas son más personales, se crea mucho ruido. A medida que el usuario indiza etiquetas más consistentes, aumenta el beneficio social.	Alta, para beneficio propio.
El etiquetado amiguista: etiquetar para compartir en un grupo reducido (amigos, compañeros, familia). Se usan etiquetas identificativas dentro del grupo pero desconocidas para otros. Ej.: Tinet.	Muy útil dentro del grupo, pero aporta poco al resto de las comunidades.	Alta, para compartir y reforzar el sentimiento de comunidad dentro de un grupo.
El etiquetado altruista: etiquetar para compartir con todo el mundo. Se escogen etiquetas generalmente comprensibles y conocidas. Ej.: música_funky.	Muy alto. Es la que más contribuye, la más generosa.	Baja. No hay un beneficio directo asociado, salvo la satisfacción personal.
El etiquetado populista: etiquetar para conseguir que algo resulte más atractivo y tenga más visitas. Ej.: Muy_interesante.	Ninguno. Es corred basura (*spam*).	Alta. Quien indiza así busca un beneficio directo y evidente.

Fuente: basado en Javier Cañada (2006).

La indización resulta barata, rápida, fácil de usar y tiene todo el espectro posible de la terminología, desde los términos más generales hasta los más específicos y actualizados (si el documento trata de Tagzania, el usuario lo indiza Tagzania sin necesidad de que un lenguaje documental controlado lo haya recogido previamente).

Ahora bien, la exhaustividad no es homogénea, ya que los objetos no se describen con el mismo grado:

- Puede haber un recurso con muchas etiquetas (exhaustividad alta) y recursos con pocas etiquetas (exhaustividad baja).

- Puede haber documentos indizados para muchas personas que nos darán enfoques diferentes sobre el mismo documento o puede haber documentos sin indizar.

Folksonomía

La indización social es el proceso distribuido en el que los recursos se describen mediante etiquetas. El resultado agregado se conoce como folksonomía [10], que significa 'clasificación hecha por el pueblo'. Son sistemas simples y eficientes. Su utilidad se deriva de la capacidad de relacionar las necesidades de los usuarios con un vocabulario habitual. No buscan la precisión.

(10) *Folksonomía*, del inglés *folksonomy*, es un neologismo. *Volk* (alemán) = 'del pueblo' + *taxis* (griego) = 'ordenación' + *nomia* (griego) = 'reglas': 'clasificación hecha por el pueblo'.

Las folksonomías tienen dos dimensiones relacionadas (Hassan Montero, 2006): la personal y la colectiva.

- En la **personal**, *personomia*, cada usuario confecciona su propio índice de etiquetas.
- En la **colectiva**, cada usuario comparte sus etiquetas y contribuye a generar un índice global de etiquetas o folksonomía. Este aspecto resulta muy interesante en indización, porque un documento descrito por cien usuarios con etiquetas coincidentes es una indización más fiable (en el sentido de recuperable) que la que haría el autor. Hassan Montero habla de indización por agregación.

Podemos clasificar las folksonomías en dos grupos (Hernández Quintana, 2008, y Weller, 2007):

- Las folksonomías estrechas o*narrow*, que son del tipo "un documento, un indizador", es decir, solo el autor puede etiquetar el contenido; sería el caso de Flickr.
- Las folksonomies generales o*broad*, en las que un documento puede ser etiquetado por varias personas, como es el caso de los marcadores sociales.

La tecnología que posibilita las folksonomías se activa el 2003 con programas como Delicious y Flickr, y tienen un aumento imparable hasta el 2006, momento en el que dichos programas ya ofrecen opciones de clusterización de las etiquetas (por ejemplo, Flickr ofrece etiquetas agrupadas por categorías). Ambos pertenecen a Yahoo.

Reflexión

En el año 2010, Yahoo, propietaria de Delicious, redactó un informe en el que anunciaba que la web llegaba a su ocaso (*sunsetted*). Muchos lo interpretaron como el cierre de la web y la comunidad social se escandalizó ante la posibilidad de perder todos los marcadores que había guardado en Delicious. La cuestión se saldó con la reventa de Delicious a la empresa Avos System. Como documentalistas, sería positivo que reflexionáramos sobre el tema y que nos diéramos cuenta de la indefensión de los usuarios ante las decisiones empresariales de productos gratuitos como este. La recomendación de los expertos es que exportemos nuestros marcadores en paralelo a otros programas, como Diigo o Mr Wong.

Miles de personas que indizan etiquetas representa un volumen considerable. Es evidente que contienen mucha información, no solo sobre el contenido del documento en cuestión, sino también sobre los propios usuarios del sistema y sus rutinas de búsqueda. ¿Qué se hace con tantas etiquetas? Básicamente, se siguen dos enfoques:

1) Aprovechar todo el conocimiento de las folksonomías para crear más conocimiento (Navoni y González, 2009):

a) Utilizar las folksonomías como complemento de otros sistemas de indización que ejerza algún control sobre las etiquetas. Se trata de aplicar técnicas de indización automática en las etiquetas, es decir, aplicar métodos estadísticos sobre frecuencia de uso y coocurrencia de las palabras.

b) Combinar las folksonomías con sistemas controlados como ontologías. Se trataría de que un lenguaje documental controlado[11] proporcionara más nombres de etiquetas, que en el mismo contexto serían útiles para la etiqueta *x* introducida por el usuario.

(11)También habría una sinergia positiva a la inversa, el lenguaje documental controlado se podría beneficiar de la aportación continua y actualizada de vocabulario, que en definitiva es lo que utiliza el usuario.

Por ejemplo, el usuario introduce la etiqueta *moneda* y la ontología le sugiere indizar, además, *bancos*, *dinero*, *acuñación*, *finanzas*, *oro*, *plata* y *riqueza*.

2) Mejorar la calidad de la indización. Se proponen dos líneas:

a) Sistemas de recomendación de etiquetas. El usuario introduce la web que quiere etiquetar y el sistema le responde con las etiquetas que otros usuarios han utilizado para la misma web, por si le resultan útiles. De este modo, se consigue cierto control sobre el vocabulario y se evitan algunos casos de sinonimia. La sugerencia es una sugerencia: el usuario siempre puede obviarla. Podemos clasificar los sitios web que permiten la indización social en dos grupos: los que permiten poner etiquetas libremente (Flickr o YouTube) y los que las sugieren (Delicious). Sugerir etiquetas beneficia la recuperación porque aumenta la coherencia entre internautas pero empobrece la espontaneidad del usuario (Marcos, 2009).

b) Alfabetizar al usuario. Son varios los autores (Hernández Quintana, 2008; Noruzzi, 2006, y Spiteri, 2007) que proponen alfabetizar al usuario dándole instrucciones para indizar. Apuntan que las folksonomías han supuesto un cambio en la metodología por la distribución y descentralización de la indización y que se podrían lograr más hitos si se organizara la forma de indizar y clasificar la información. Algunos de los puntos que se proponen son la redacción de normas sobre:

- el uso de sustantivos cuantitativos y no cuantitativos;

- la elaboración de etiquetas compuestas (por ejemplo, con un espacio o guión entre unitérminos);

- la evaluación de la calidad o aplicaciones de cada ítem;

- el uso de enlaces a diccionarios que actúen como autoridades y controlen la forma de la etiqueta;

- el añadido de nombres personales provenientes de listas de autoridades y del rol que tiene con el concepto que se etiqueta;

- el añadido de todo tipo de facetas (*faceted tagging*): geográficas (nombres geográficos provenientes de lenguajes controlados como tesauros), de tiempo, de forma, de género.

Las propuestas que hacen referencia a copiar la etiqueta desde un vocabulario controlado (diccionario, tesauro o clasificación) son las más interesantes, y hay bastantes artículos que proponen usar la LCSH, la CDU o tesauros, pero también se propone indizar a partir de los artículos de la Wikipedia (creados de manera colaborativa y con el mismo espíritu intuitivo de las etiquetas) como vocabulario controlado.

Observación

Fijaos en que si el internauta elige un término sugerido, venga de la Wikipedia, del WordNet o de un cálculo estadístico del Delicious, ya está indizando de manera controlada y no libre. Con todo, el cambio no estriba en la tipología libre respecto de la controlada, sino en una tipología nueva, lo que en inglés se denomina ***user vocabulary*** (o proveniente de la colaboración social), ante el ***controled vocabulary*** (vocabulario hecho por profesionales).

La recuperación con descriptores libres

La indización con descriptores libres, que todo el mundo ha hecho de manera individual (persona que indiza su biblioteca personal), adquiere una nueva dimensión cuando miles de personas hacen lo mismo. A pesar de los inconvenientes de la falta de control sobre el vocabulario, que son evidentes, es tan grande su aportación en el mundo de la Web que, a pesar de ser imperfecta, resulta muy útil en la recuperación.

Ventajas e inconvenientes de la recuperación con descriptores libres.

Ventajas	Inconvenientes
1) La comunidad se beneficia de un volumen ingente de documentación medianamente descrita. Su calidad puede ser discutible, pero está operativa, es accesible. **2)** Se rompe la subjetividad de un único indizador. **3)** Los puntos de acceso son más diversos. **4)** No necesita traducción de los conceptos del lenguaje natural de los documentos a un lenguaje artificial. **5)** Se trata de un tipo de lenguaje rápido y fácil de actualizar. **6)** Se adapta perfectamente a los usuarios y tipos de SID, ya que es un lenguaje hecho a medida. **7)** No hace falta una formación previa de los analistas. Precisamente la ausencia de reglas y principios hacen innecesaria la formación. **8)** Indizan texto pero también imagen fija (foto) y en movimiento (vídeo, película). **9)** El vocabulario presenta una autoridad de usuario. **10)** El número de indizadores aumenta la tasa de consistencia.	**1)** Todos los que se derivan del lenguaje natural: • Sinónimos. • Polisémicos. • Falta de términos relacionados que amplíen la búsqueda. • Siglas o acrónimos. • Palabras sin significado en determinados contextos (ej.: la palabra *tuya*, que solo tiene significado en Botánica). **2)** *Ego-centered tag* o etiquetas con términos vacíos para la comunidad, puesto que solo tienen sentido individualmente. **3)** Nivel de exhaustividad diverso, no todos los documentos están indizados con el mismo grado.

En resumen:

La indización social participa en las características de las listas de descriptores libres en la filosofía de la indización, ya que cada participante indiza unos descriptores libres seleccionados según un proceso intelectual a partir del examen del recurso sin verificar si los descriptores propuestos existen o no en un lenguaje controlado. A medida que han transcurrido los años, el volumen de etiquetas ha permitido ir más allá y crear un vocabulario de términos con autoridad de usuario (*user vocabulary*). Sobre sus términos se pueden efectuar cálculos estadísticos y seleccionar las etiquetas con la tasa de coherencia entre indizadores más elevada o hacer clusterización. El paso siguiente será importar las etiquetas de otros lenguajes, esta vez controlados, como listas de autoridades (para los nombres propios), tesauros (para nombres geográficos), etc. La Web semántica permite a los descriptores libres crear sistemas basados en lenguaje natural y libre, que poco a poco se irán estructurando y controlando. La meta es una Web semántica con ontologías.

2.3.5. Indización automática

La indización automática es el método por el cual un ordenador aplica un algoritmo (o programa) a un documento electrónico para identificar los términos que puedan representar la materia y ser utilizados como términos de indización y recuperación en un índice o lista.

La indización automática es, junto con la social, la alternativa más viable para indizar la Web.

Indización automática

La indización automática es un lenguaje natural, libre, poscoordinado, alfabético y analítico para palabras clave.

¿Cómo se indiza automáticamente?

El primer paso es leer el texto. Para hacerlo, es preciso que el documento se encuentre en formato electrónico y sea accesible. Esta afirmación tan sencilla implica:

- excluir la documentación audiovisual, imagen fija (fotografías) o en movimiento (vídeo) que habitualmente no va acompañada de texto;

- excluir también toda la documentación que pertenezca a intranets (donde hace falta contraseña) y toda la que se genere de forma dinámica (contenida en bases de datos), lo que conocemos como *Internet invisible* y que se calcula que supera en cinco veces la Web visible.

Después se toma una serie de decisiones.

1) El documento electrónico puede ser un texto plano con algún campo tipo *resumen* y *palabras clave* o puede estar estructurado con metadatos, tanto para el contenido como para la forma. **Hay que decidir si el programa se aplicará en el texto completo o en campos determinados del documento**; por ejemplo, solo en el campo *palabras clave*. La calidad del resultado será muy diferente en un caso o en otro: en el primer supuesto, será el programa el que seleccionará las palabras más representativas –por repetidas– del texto con cálculos estadísticos, mientras que en el segundo, los términos de indización ya han sido seleccionados mediante un proceso intelectual.

Recordemos que los metadatos son datos formalizados y que suponen una pieza clave de la Web semántica, junto con el lenguaje XML y el formato RDF.

Observación

El XML es un lenguaje que presenta las propiedades del HTML y la posibilidad de incluir en el nivel de código una infraestructura de metadatos que explicite la información del recurso.

2) ¿Qué hay que hacer con los términos que contienen números, signos de puntuación, guiones, mayúsculas/minúsculas y acentos? Por lo general, se trata de caracteres que no aportan significado, pero que en determinados contextos pueden ser determinantes.

Número: N2, TV1.

Puntos, guiones, signos (www.uoc.edu), Fuentes_Información (es una etiqueta propia de Delicious).

Acentos (útiles para diferenciar diacríticos): en catalán, os/ós; en castellano, te/té.

Observación

El RDF es un marco de descripción de recursos (*resource description framework*, *RDF*) para metadatos desarrollado por el World Wide Web Consortium (W3C).

3) ¿Qué hay que hacer con las palabras vacías (artículos, pronombres, preposiciones, conjunciones, adverbios, numerales)? Son palabras muy frecuentes, pero que aportan poco valor de contenido. Se conocen como listas de detención en español y *stopword list* en inglés. Los programas de indización automática tienen un fichero con las palabras vacías que hay que obviar. Ahora bien, este fichero puede estar implementado de tres maneras diferentes:

Ejemplo de metadatos

Haced clic en el icono *Indización* de la base de datos de revistas de la Universidad de Murcia: http://revistas.um.es.

a) Predeterminado. Desde el principio el sistema dispone de la lista de detención en su idioma o idiomas. De hecho, su realización es fácil, puesto que solo hay que añadir las categorías vacías de una base de datos de terminología en el idioma deseado. Los artículos y las conjunciones siempre son los mismos, incluso los verbos se pueden llegar a contabilizar y flexionar en todos los tiempos verbales.

b) Contextualizado (*stop word context-dependent*). Cada sistema elabora la lista de detención según su ámbito temático. Contextualizar la lista permite evitar dos graves inconvenientes:

- Palabras con significado que se vuelven vacías.

En un centro especializado en medicina del deporte todos los documentos harán referencia a *medicina del deporte* y, por lo tanto, dicha palabra estará vacía en ese contexto.

- Palabras vacías que se vuelve importantes en la indización.

En un texto de historia, los números (1319-1387), numerales (Pere III) y los adjetivos pueden tener una gran carga significativa (el Ceremonioso). En este ejemplo podemos observar que *Pere III el Ceremonioso 1319-1387* podría quedar indizado como *Pere* si no se mantienen algunas palabras vacías.

c) Evitado expresamente para permitir al sistema la búsqueda por frases y sintagmas.

Por ejemplo, para recuperar un concepto como el nombre del diario *El País*, en el que el artículo tiene un papel importante. Los sistemas que los evitan disponen de otras herramientas para reducir significativamente el número de palabras indizadas, como técnicas de *stemming* o lematización. En este sentido, más adelante hablaremos de los marcadores discursivos, en los que veremos como palabras en principio vacías ayudan en gran medida en la decisión de qué términos seleccionar.

4) Aplicar métodos estadísticos. Una vez eliminadas las palabras vacías, nos queda un conjunto de unitérminos con significado, pero aun así su número puede ser muy elevado. El paso siguiente consiste en seleccionar las de mayor relevancia en la descripción del documento. Este paso se resuelve aplicando varios métodos estadísticos (o lingüísticos y semánticos, que veremos más adelante), bien en un orden secuencial, bien alternando los métodos.

Los métodos estadísticos han sido la primera aproximación a la indización automática y todavía hoy en día son una parte consustancial de ella. La teoría de fondo es el cálculo del peso (ponderación) de las palabras: ni las palabras más repetidas (por vacías) ni las menos repetidas (por específicas) son adecuadas para ser seleccionadas. Los métodos estadísticos aplicados en PLN son de tres tipos (se pueden usar solos o combinados):

PLN

El procesamiento del lenguaje natural (PLN o *NLP*, de su nombre en inglés, *natural language processing*) es la disciplina informática que se encarga de tratar computacionalmente las lenguas naturales o lenguajes humanos.

En la actualidad, las principales aplicaciones o áreas de trabajo del PLN son las siguientes:

- recuperación de la información,
- extracción de la información,
- búsqueda de respuestas,
- traducción automática,
- generación de resúmenes, y
- reconocimiento del habla.

a) Frecuencia. Hans Meter Luhn (1957) aplica la ley de Zipf al campo de la indización automática. Luhn propone los pasos siguientes: calcular la frecuencia de todas las palabras del texto o colección, ordenarlas en orden decreciente, eliminar las de frecuencia más alta, eliminar las de frecuencia más baja e indizar con el resto.

b) Frecuencia inversa. Sparck Jones (1972) puso de manifiesto la capacidad de discriminación de un término frente a otro. Esta discriminación se tiene que considerar en el conjunto de la colección, no en un único documento. Hay que comparar las palabras clave entre los documentos del fondo para detectar cuáles son realmente discriminativas.

c) Discriminación. G. Salton (1989), a partir de la idea de que las palabras de un texto se clasifican según su capacidad para discriminar unos documentos de otros en una colección, ideó un sistema de indización conocido como el **modelo de valor de discriminación**, que atribuye el peso o valor más alto a aquellos términos que causan la máxima separación posible entre los documentos de una colección. Es decir, el valor de un término depende de cómo varía la separación media entre los documentos. Por lo tanto, las mejores palabras son las que consiguen la mayor distancia. El análisis del **valor de discriminación** asigna una función específica en el análisis de contenido a las palabras simples, a las yuxtapuestas, a las frases y a grupos de palabras.

5) Métodos lingüísticos. Los primeros analizadores lingüísticos datan de las décadas de 1960 y 1970. Su aportación al análisis del contenido resulta capital, ya que permiten analizar el texto en tres niveles de profundidad: palabra, frase y texto.

Cada uno de estos niveles es analizado por módulos del programa basados en diferentes disciplinas:

Palabra	Morfología
Palabra dentro de la frase	Sintaxis
Palabra dentro del texto	Semántica

Con estas operaciones se consigue un fichero inverso en el que constan los unitérminos y los documentos en que aparecen. Cada unitérmino va asociado a un documento y a una posición dentro del documento (por ejemplo, al título).

6) Métodos semánticos. La semántica es la ciencia que estudia el significado de las palabras. Es una pieza fundamental dentro del PLN y la Web semántica, valga la redundancia. Algunas de las propuestas son los marcadores discursivos y la participación de lenguajes controlados en tareas de indización automática.

a) Los marcadores discursivos

El PLN todavía está lejos de ofrecer sistemas capaces de entender semánticamente un texto, como lo haría una persona, pero está trabajando en una línea muy interesante, que son los marcadores discursivos. Se trata de dotar al algoritmo del robot de las relaciones semánticas que se derivan de cinco grupos de marcadores y, a partir de aquí, inferir un conocimiento.

Los marcadores discursivos son unidades lingüísticas invariables, por lo cual son automatizables. Los cinco grandes grupos son los marcadores (Portolés).

Lectura complementaria

Para más información sobre cada marcador discursivo, podéis consultar el *Diccionario de partículas discursivas del español*, de Briz, Pons y Portolés (http://textodigital.com/p/ddpd/).

Ejemplos de algunos marcadores discursivos.

Marcadores	Ejemplos
Estructuradores de la información	Primero, segundo. Por un lado, por otro. Después, entonces.
Conectores	Incluso, es más. Así pues, por lo tanto. Aun así, sin embargo.
Reformuladores	Es decir, a saber, en otras términos. En todo caso, en cualquier caso.
Operadores argumentadores	En realidad, en el fondo. En concreto, en particular.
Marcadores conversacionales	Naturalmente, sin duda. ¿Verdad? ¿Eh?

Uno de los marcadores estructuradores son los marcadores ordenadores que agrupan varios ítems como si fueran partes de uno solo, como por ejemplo:

- numéricamente: primero, segundo, etc.;
- en el espacio: por un lado, por otro;
- en el tiempo: después, entonces, en fin.

Si el programa dispone de estos marcadores, podrá inferir un discurso más elaborado a partir del documento y controlará mejor las partes discursivas (introducción, cuerpo, conclusiones) y las partes orgánicas del texto.

El programa mantendrá unido el conjunto de ítems que, de una forma u otra, estaban ordenados con los marcadores anteriores.

Así, si el texto decía "primero Namibia, segundo Venezuela, tercero Nepal...", el programa indizará los tres nombres y no solo uno, y los mantendrá relacionados.

Si el texto decía "[...] lo que investigaba en el fondo era el sodio", el programa detectará un marcador argumentador (*en el fondo*) e indizará la primera palabra con significado que vaya detrás (*sodio*).

Observación

Fijaos en que cualquiera de estos marcadores discursivos se podría haber catalogado como una palabra vacía, ya que son adjetivos, conjunciones y adverbios, y el programa habría perdido una información muy valiosa a la hora de mantener indizadas partes del texto.

b) La participación de lenguaje documental controlado

Se trata de una indización semiautomática, a diferencia de las anteriores, completamente automáticas.

A grandes rasgos, el funcionamiento consiste en el hecho de que el robot detecta las palabras más significativas del documento y las compara con un vocabulario controlado, como un tesauro o algún tipo de clasificación, que propone un término controlado para indizar a partir de sus referencias.

En algunos sistemas este último paso es automático, mientras que en otros es una persona quien valida la decisión. Los sistemas semiautomáticos de categorización pueden ser de tres tipos:

- Categorización basada en reglas.
- Categorización basada en el aprendizaje automático a partir de documentos ejemplares.
- Una combinación de los dos modelos anteriores. Es la opción que mejores resultados da, pero hay que dedicar un tiempo al diseño de las reglas y al entrenamiento de documentos ejemplares.

7) La indización automática no es solo una manera de indizar y, por lo tanto, un lenguaje documental en sí, sino que **también es una aplicación** de la que se benefician todos los lenguajes documentales.

A lo largo de cada lenguaje, se ha tratado la forma en que la automatización de los procesos de indización y recuperación puede agilizar todo el proceso. Así, hemos visto cómo se puede clasificar de manera automática o semiautomática, cómo se puede descomponer un encabezamiento de materia controlado en una sucesión de palabras clave, cómo se pueden crear tesauros o indizar con un tesauro de manera automatizada, el papel relevante de las etiquetas y los cálculos estadísticos que se pueden ejecutar para sugerir nuevas etiquetas.

De cara al futuro, lo más interesante es ver la forma en que los lenguajes documentales más potentes y más experimentados se mantienen al día de la Web semántica, y ya los tenemos en formato SKOS:

- Ex CDU en SKOS (http://www.udcc.org/udcsummary/exports.htm),
- LCSH en SKOS (http://id.loc.gov/techcenter/metadata.html),

- la clasificación Dewey (http://oclc.org/developer/documentation/dewey-web-services/using-api).

La recuperación de información indizada automáticamente

Buscadores

En la Web se puede buscar de dos maneras: **navegando** o con **buscadores**. Es decir, podemos llegar a encontrar un dato saltando de una página a otra a partir de los enlaces o bien escribiendo los términos que queremos en la caja de un buscador. El primer sistema no implica ninguna tarea de indización; el segundo, sí, y es una indización automática.

Ejemplo

El autor de un blog cuelga un apunte sobre sus vacaciones en Sicilia. El autor no ha indizado el contenido del artículo, pero nosotros podemos llegar a él ya sea saltando de una página que tenía enlazada, ya sea buscando en Google.

Los **algoritmos de los buscadores** comparan la palabra de la búsqueda con las palabras contenidas en los textos de su base de datos. Funciona bien para textos, pero no para material gráfico y audiovisual que no incluya texto o pie de fotografía.

El usuario tiene la sensación de que el buscador rastrea toda la Web buscando los términos que ha pedido como si fuera en tiempo real, pero esto es una ilusión, porque sería mecánicamente imposible (miles de usuarios buscando en paralelo en Google y recibiendo respuestas en tiempo real). En realidad, los buscadores no rastrean la Web en el momento de la consulta, sino en el momento de la indización. Rastrean y crean sus ficheros inversos, que se van actualizando.

Cuando el usuario lleva a cabo una búsqueda, el programa no consulta la Web, sino su base de datos del fichero inverso, por eso se obtiene el resultado en cuestión de segundos.

La indización automática no plantea grandes problemas, salvo uno, que es el orden en que se presentan los miles de resultados que se encuentran. Las soluciones han ido evolucionando en el tiempo: primero eran los documentos que contenían los términos, después las búsquedas acotadas con los operadores booleanos, más tarde Google introduce el concepto de relevancia de la fuente en función de los enlaces que tiene y que recibe, es decir, ya no solo se considera la calidad interna de la fuente, sino también la calidad externa que le atribuyen otras fuentes.

Observación

Fijaos en que es el mismo criterio de evaluación de la calidad que se utiliza con las publicaciones periódicas y el factor de impacto, como el JCR de ISI web of knowledge, In-recs, RESH, etc.

Estadísticas de buscadores

Los tres buscadores más utilizados según las estadísticas son, por orden, Google, Yahoo y Bing (AOL lo es en América).

Recuperación en una web estructurada

La recuperación tal como la entendemos hoy en día sufrirá una **revolución** por el uso de ontologías y los motores de inferencia.

El futuro se presenta más enfocado hacia las buscas en contexto más apropiadas para estos nuevos usuarios-editores-documentalistas. Se pretende utilizar los metadatos para efectuar cálculos sobre la relevancia de la Web, para la navegación por facetas (por lugar, tiempo, forma o cualquier otra faceta propia de un tema) y para buscar por fórmulas que otros usuarios hayan empleado reiteradamente.

Como afirma Mendez citando a Witten, Gori y Numerico, nos dirigimos hacia una "diversidad descentralizada", en la que interrogaremos a la Web de varias maneras y en la que coexisten con una anarquía organizada los datos entrelazados (documentos, opiniones, relaciones, etc.).

Una de las ventajas de los metadatos, es decir, de partir de documentos estructurados, es que el usuario podrá buscar en la Web como busca en una base de datos, **por campos**.

Esto significará que podrá acotar la búsqueda, por ejemplo, pidiendo documentos en los que se hable de *Bedrich Smetana* como tema y no recuperar toda la obra de este músico (equivaldría a un catálogo pedir *Bedrich Smetana* como materia o *Bedrich Smetana* como autor).

Otra aplicación son los **sistemas de búsqueda de respuestas**, que responderán directamente a la pregunta, no ofrecerán un conjunto de documentos en los que aparezca el término de la consulta, sino que aparecerá directamente el fragmento con la respuesta.

Desde el punto de vista de la recuperación y lenguajes documentales, son interesantes dos técnicas de esta "diversidad descentralizada": los vocabularios poscontrolados y las técnicas de clusterización. Las dos técnicas parten de un vocabulario libre que el programa acabará por controlar.

1) Los vocabularios poscontrolados (Lancaster). Se constata que los usuarios hacen búsquedas cortas de uno o dos términos, que vuelcan muchos resultados. El usuario no hace búsquedas largas y elaboradas con operadores booleanos, pero los buscadores pueden almacenar las búsquedas de otros usuarios y sugerir al usuario que busque por ese concepto y otro más. De alguna manera, el buscador está indizando la pregunta y guarda la fórmula para otros usuarios. El vocabulario es libre pero el robot lo controla.

Ejemplo

Los usuarios acostumbran a pedir *monovolúmenes*, pero el programa ha almacenado la fórmula (Monovolúmenes) and (Seat or Volkswagen or Nissan...), que recuperará de forma más exhaustiva. De hecho, el programa está recogiendo los TE y TR (términos específicos y términos relacionados) de monovolúmenes.

2) Sistemas de clústeres. La clusterización de datos es una técnica muy común en el análisis estadístico de datos. Básicamente, se trata de la clasificación de objetos similares en diferentes grupos. Los clústeres son carpetas clasificadas

según la coaparición de los términos en el texto. Se supone que cuanto más a menudo aparezcan juntos los términos de un tema determinado, más probable será que sus significados estén relacionados. El programa presenta las carpetas o los clústeres en que aparece el tema que se busca, de esa forma el usuario puede escoger el enfoque que más le interese.

Ejemplo

Un usuario busca el término *lista de palabras vacías* en el buscador Yippy (http://search.yippy.com) y este da noventa registros clasificados en diez carpetas iniciales (algunas carpetas se abren) para que el usuario escoja: Search, My SQL Manual, Tools, Download, etc. En este caso, el programa ha sintetizado el contenido de los resultados en forma de taxonomía.

Web semántica: indización y recuperación

La Web semántica es un conjunto de iniciativas destinadas a promover una futura Web con páginas organizadas, estructuradas y codificadas de tal manera que los ordenadores sean capaces de efectuar inferencias y razonar a partir de sus contenidos.

Será una **gran base de datos** capaz de soportar un procesamiento sistemático y coherente de la información (Codina y Pedraza, 2007).

La Web semántica se basa en un lenguaje XML y en unos formatos comunes (RDF) que permiten la interoperatibilidad (*linked data*) con independencia de la plataforma desde la que se trabaje.

La indización en la Web semántica se fundamentará en la información estructurada: los recursos web estarán descritos -es decir, indizados- en forma y contenido con metadatos (que pueden haber sido generados de forma manual o automática), se buscará con agentes inteligentes que se adaptarán a nuestra situación y los términos de indización se interrelacionarán a partir de ontologías.

Parece que lo más sensato es pensar que la indización en la Web semántica consistirá en una combinación de todos los sistemas actuales, así:

- Se seguirán indizando de manera intelectual con lenguajes controlados (clasificaciones, encabezamientos de materia, autoridades y tesauros) las fuentes de información lo bastante valiosas para que el resultado no esté condicionado por la inversión económica, como por ejemplo las bases de datos de artículos en ciencias de la salud, como MESH.

- El uso de vocabularios controlados altamente formalizados y un PLN cada vez más potente propiciarán la implementación de ontologías. Se crearán

ontologías de forma automática y manual, y se indizará automática y manualmente a partir de ontologías.

- Se indizará de manera semiautomática o semiasistida la gran mayoría de la Web, que por sus dimensiones no permite otras posibilidades. Y se espera que cada vez más los documentos electrónicos vengan de serie con metadatos. Tales metadatos, a su vez, pueden haber sido generados de manera intelectual o por un robot automático.

- Se indizará socialmente con lenguajes libres como los descriptores libres o etiquetas, sobre todo la información audiovisual que no es fácilmente indizable de manera automática por no incluir texto. En este sentido, se está investigando en robots que reconozcan formas simples en las imágenes; de todos modos, hasta que no sean una realidad, la mejor opción son las etiquetas de los internautas.

Un caso interesante: los wikis y las ontologías

Podemos encontrar dos enfoques: el primer enfoque, que considera un wiki como una ontología en la que las páginas son tratadas como conceptos y los enlaces que aparecen en ellas se consideran relaciones. A medida que se crea el wiki, se crea la ontología. Y el segundo enfoque, que parte de la existencia previa de una ontología a partir de la cual etiqueta semánticamente las páginas y relaciones del wiki.

La recuperación en la Web semántica consistirá, como indica Berners-Lee, no en una inteligencia artificial mágica que permita a los ordenadores entender las palabras de los usuarios, sino en la habilidad de una máquina para resolver problemas bien definidos a partir de operaciones muy definidas que se llevarán a cabo sobre datos muy definidos (W3C, 1999).

Webs recomendadas

Buscador en la Web semántica http://swoogle.umbc.edu

Sobre metadatos: http://ca.wikipedia.org/wiki/metadades

3. Calidad y coherencia en la representación de contenidos

La **calidad** y la **coherencia** de la indización dependen de factores como la competencia del indizador y la calidad de los instrumentos o lenguajes documentales. La coherencia es un factor importante en el comportamiento de un sistema de indización, especialmente cuando forma parte de una red de centros y la información se tiene que intercambiar entre ellos.

La coherencia se calcula de la siguiente manera: dos analistas indizan el mismo documento, con un lenguaje de descriptores como un tesauro. Se cuentan separadamente el número de descriptores idénticos entre los dos analistas sobre el total de descriptores.

Como ejemplifica van Slype:

- El documentalista 1 ha asignado los descriptores A, B, C, D, E, F.
- El documentalista 2 ha asignado los descriptores A, C, D, F, G, H.
- Hay 4 descriptores idénticos A, C, D, F y un total de 8 descriptores diferentes. Tasa de coherencia = 4/8 = 50% (van Slype, 1991, p. 123).

La consistencia en la indización suele oscilar entre el 20% de mínima y el 60% de máxima (Isidoro Gil, 2001).

A modo de conclusión

La norma UNE 50-121-91 *Métodos para el análisis de documentos, determinación de su contenido y selección de términos de indización* establece tres fases:

- Examinar el documento para identificar su contenido: el analista tiene que examinar con precisión el documento. La lectura completa es a menudo impracticable, pero sí que tiene que prestar atención al título, resumen, sumario, introducción, ilustraciones y palabras o frases destacadas en una tipografía diferente.

- Seleccionar los conceptos principales de los contenidos: el analista tiene que identificar las nociones que son elementos esenciales de la descripción del contenido, tiene que ser consciente del número de conceptos (criterio de exhaustividad) y la exactitud de los mismos (criterio de especificidad).

- Traducir a un lenguaje documental: para traducir el concepto inicial escrito en lenguaje natural a un lenguaje documental hay que consultar el listado del lenguaje buscando la forma aceptada.

Lectures complementaries

Podéis ampliar la información sobre la coherencia en la indización leyendo las obras siguientes:

G. van Slype (1991). *Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales*. Madrid: Pirámide. Fundación Germán Sánchez Ruipérez. Biblioteca del Libro.

I. Gil Leiva (2008). *Manual de indización. Teoría y práctica*. Gijón: Ediciones Trea.

3.1. La calidad del indizador

En este apartado analizaremos el papel que tenemos nosotros como indizadores. Sin embargo, antes hagamos una lista de las ventajas que nos facilitan la tarea:

1) Hay temas más fáciles de indizar que otros por el conocimiento que tenemos de ellos.

2) Algunos lenguajes son más fáciles que otros, como los poscoordinados, que nos ahorran conocer las reglas de precoordinación. Los "fáciles" son las listas de autoridades, los tesauros, los descriptores libres y la lista de palabras clave.

3) Es más fácil indizar un dato que una materia.

Es más fácil indizar *Aristóteles* que las materias de algunas de sus obras. Para indizar *Aristóteles*, solo hay que consultar una lista de autoridades como Lenoti, y un *véase* propio de una relación de equivalencia nos dice que tenemos que indizar *Aristóteles, 384-322 aC*. Haciendo doble clic tenemos la autoridad aceptada. En cambio, indizar la materia de una de sus obras resulta más laborioso (decidir qué rama de la filosofía, conceptos, etc.).

4) Si el SID dispone de manuales y tutoriales sobre el grado de exhaustividad y especificidad que quieren, nos sentiremos más guiados.

Aun así, como indizadores podemos cometer dos tipos de errores: los **técnicos** y los **éticos**.

3.1.1. Errores técnicos

Partimos del supuesto de que el indizador no cometerá errores de conocimiento del lenguaje que tiene entre manos, como por ejemplo no entender las referencias de *véase* de un término no aceptado a uno aceptado. Con todo, puede cometer los errores que indicamos a continuación.

1) Errores en la **selección del tema del documento**: el indizador no ha captado la verdadera materia del documento. Las causas pueden ser la falta de atención o el desconocimiento de la materia.

2) Errores en la **selección numérica de los términos**: se puede equivocar obviando temas interesantes, es decir, por ejemplo, el documento trata de cuatro temas pero solo escoge dos.

3) Errores en la **selección del término**: el indizador elige un término más genérico de lo que sería deseable por una falta de especificidad del lenguaje documental. La ausencia del término lo obliga a indizar con un término conceptualmente más genérico. El indizador comprende la materia, pero el lenguaje documental no le permite expresarse.

4) Errores por **omisión**: el documento trata de un tema que no aparece en el lenguaje documental y, ante la duda de lo que es, no lo indiza.

El documento trata sobre las aplicaciones de la Apple Store y, dado que no aparece en el lenguaje, no indiza nada, cuando lo mejor sería indizar un término genérico como *comercio electrónico*. Según Lancaster, un diez por ciento de los errores en la exhaustividad se debe a omisiones. Se solucionarían si el lenguaje dispusiera de referencias de términos equivalentes y de términos relacionados, tipo *Apple Store*, *Microsoft Store TR Comercio electrónico*.

5) Errores en la **formalización**: se equivoca en la grafía del término.

351.8w, por 351.82 (Administración pública de la economía en la CDU), o Dietnes por Dientes. Este error se soluciona no tecleando el término sino copiándolo de ficheros o listas de autoridades.

6) Errores en la **coherencia al equivocarse con la sintaxis del lenguaje precoordinado**, hecho que impide reunir todos los documentos que tratan del mismo tema.

La falta de consistencia se puede dar en varios niveles, que ejemplificaremos a partir de un encabezamiento compuesto, como es Dientes – Cuidado e higiene – Estadísticas.

En el caso óptimo de que todos los indizadores conozcan la precoordinación, encontraremos ordenados todos los documentos indizados por la secuencia Dientes, como por ejemplo:

Dientes – Cuidado e higiene – Estadísticas.

En cambio, si un indizador altera el orden de los subencabezamientos, se producirá una mezcla en la que perderemos documentos.

Cuidado e higiene – Dientes – Estadísticas

Si un indizador indiza con un término genérico de*dientes*, también perderemos la secuencia

Boca – Cuidado e higiene – Estadísticas.

7) Errores en el**almacenamiento en el catálogo**: son errores técnicos derivados del programa de gestión (falta de espacio en los campos, en la memoria, etc.).

Los dos primeros errores no guardan relación con el vocabulario del lenguaje. Los siguientes sí, y es en estos últimos casos en los que un lenguaje muy construido puede ayudar a minimizarlos: con términos genéricos abiertos en suficientes términos específicos, términos no utilizados que remiten con *véase* a los términos usados, con notas de aplicación y notas explicativas a los descriptores, con referencias cruzadas y términos relacionados. Cuanto más rico sea el lenguaje, menos conocimientos en la materia debe tener el indizador.

3.1.2. Errores éticos

Hay tres tipos de errores éticos: los de discriminación (u ofensa), los de censura y los intencionados.

1) Errores por discriminación u ofensa

Hay que evitar términos que puedan resultar ofensivos o discriminatorios por cuestiones de género, raza, religión, condición, etc.

El control del vocabulario es una herramienta de gran valor en este cometido, ya que los lenguajes controlados han pasado por una criba de conceptos en la que la mayoría de los términos ofensivos han sido rechazados. Y decimos la mayoría porque algunos lenguajes todavía arrastran concepciones antiguas que cuesta modificar. En la bibliografía científica sobre encabezamientos de

materia encontramos muchos artículos que analizan temas sensibles comparando encabezamientos en dos listas y que piden una revisión urgente de los epígrafes.

Lectura recomendada

Para ampliar este tema, recomendamos la lectura de Carmen Caro y R. San Segundo, *Lenguajes documentales y exclusión social* (http://dialnet.unirioja.es/servlet/articulo?codigo=1300420), donde se analizan encabezamientos que ponen bajo el mismo término genérico a las madres solteras y a los delincuentes dentro del grupo de marginados sociales, o que relacionan dos términos tan dispares como *anarquismo* e *idiotez*. Los sistemas de clasificación también cometen errores éticos al mantener, por ejemplo, la rúbrica de la clase 159.922.76 para niños con defectos físicos, mentales y superdotados.

En estos casos, es recomendable no emplear tales términos para indizar y proponer un acuerdo interno del SID para sustituirlos. Si indizamos con un lenguaje en línea, accederemos a todas las actualizaciones, pero en el caso de que nuestro lenguaje esté en papel, habrá que comprobar en las actualizaciones de la web si el término ofensivo ya ha sido modificado o no.

Ejemplo

Por ejemplo, el consorcio de la CDU vela por el mantenimiento y la actualización del Master Reference File, y en esta dirección, http://www.udcc.org/major_changes.htm, podemos comprobar el estado del término que nos (pre)ocupa.

En entornos de indización libre, como buscadores generales o marcadores sociales, podemos encontrar etiquetas sobre temas sensibles expresados de manera vejatoria o sectaria, puesto que nadie más que el propio autor del texto o el internauta toma la decisión de indizarlos.

2) Errores por censura

Todas las fases de la indización están influidas por cierto grado de subjetividad del analista (por su formación, convicciones políticas, creencias religiosas, etc.), pero el documentalista, tal como recoge el código ético de la American Library Association, debe distinguir entre sus convicciones personales y sus responsabilidades profesionales y no permitir que las creencias personales interfieran en la representación del contenido de los documentos.

> "[...] We distinguish between our personal convictions and professional duties and do not allow our personal beliefs to interfere with fair representation of the aims of our institutions or the provision of access to their information resources [...]."
>
> *Code of Ethics of the American Library Association*: http://www.ala.org/advocacy/proethics/codeofethics/codeethics

3) Errores intencionados

Un tercer tipo de error ético es indizar intencionadamente de manera equivocada para conseguir una ganancia, como por ejemplo un mejor posicionamiento web. Esto se conoce como falseamiento de índices o *spamdexing*. Consiste en indizar conceptos que nos aseguran más visibilidad en la Red (por ejemplo, *muy interesante*) aumentando las referencias cruzadas y enriqueciendo los enlaces hacia la página web. Para evitar el falseamiento de índices o para comprobar que las etiquetas que hemos asignado a una web no se consideren falseadas, vale la pena consultar antes las políticas de los buscadores.

Web recomendada

Herramientas para administradores de webs (*webmasters*) de Google: http://support.google.com/webmasters/bin/answer.py?hl=se&answer=35769

3.1.3. ¿Cómo se mide la calidad de un indizador?

La calidad de un indizador se mide en comparación con otro. Esta operación se resuelve calculando la tasa de coherencia.

Partiremos de un caso delimitado: dos documentalistas, diez documentos y tres descriptores. La fórmula de la tasa de consistencia es:

c / a + b – c

Leyenda:

***a* equivale** a términos indizados en *Indizador a*.

b equivale a términos indizados en*Indizador b*.

c equivale a términos comunes en las dos indizaciones.

Descriptor	Docu-men-talista	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8	Doc 9	Doc 10
Cadaqués	A		x	x	x		x	x	x	x	x
	B		x		x		x			x	
Parque natural	A	x	x	x	x						
	B	x	x	x							
Cala Culip	A	x				x			x		
	B			x			x				

Descriptor Cadaqués: 4/8 = 50%.
Descriptor Parque natural: 3/4 = 75%
Descriptor Cala Culip: 0/5 = 0%

Nota

Algunas características de los lenguajes favorecen o dificultan la coherencia. La CDU ha sustituido el uso del signo subdividir por el *colon* (:) y otras facetaciones (como en la tabla 9) porque los indizadores interpretaban mal las instrucciones y daban lugar a tasas de coherencia muy bajas.

3.2. Evaluación de la recuperación

En la recuperación se evalúan conceptos de **microevaluación** (silencio, ruido) y de **macroevaluación** (exhaustividad y precisión). Comparándolos alcanzamos el concepto de consistencia o coherencia, que ya hemos visto anteriormente.

Partimos del cuadro siguiente (Lancaster y Van Slype), en el que podemos observar todas las posibilidades que se producen en la recuperación:

Leyenda de los elementos de la recuperación

	Pertinentes	No pertinentes	Total
Extraídos	A (aciertos)	B (ruido)	A + B (recuperados)
No extraídos	C (pérdidas)	D (correctamente rechazados)	C + D (no recuperados)
Total	A + C (total de documentos relevantes)	B + D (total de documentos no relevantes)	A + B + C + D (colección entera)

¿Cómo se calculan los documentos pertinentes y no pertinentes? Es preciso que el usuario valore como pertinente o no pertinente el conjunto de documentos que el sistema le ha dado. De los cuatro valores (A, B, C, D), podemos saber A porque son los que se han recuperado y el usuario considera relevantes y B, porque no los considera relevantes. En cambio, para saber C y D necesitaríamos un entorno ideal donde el usuario pudiera ver toda la colección y decidiera cuáles habrían sido pérdidas y cuáles no. Dado que no podemos hacerlo debido al volumen de la colección, se toma una sección y se extrapola el resultado.

Este ejemplo servirá para argumentar el resto del módulo: imaginemos que hemos buscado por documentos que contengan el término *Cadaqués*:

	Pertinentes	No pertinentes	Total
Extraídos	5	2	7
No extraídos	3	30	33
Total	8	32	40

3.2.1. Microevaluación: silencio y ruido

Tasa de silencio: c / a + c.

En la búsqueda sobre Cadaqués observamos que es 3 / 5 + 3 = 0,375, es decir, el 37,5% de los documentos pertinentes no se ha recuperado. La tasa de silencio es del 37,5%.

Tasa de ruido: b / a + b

Sobre el mismo ejemplo, es 2 / 5 + 2 = 0,285. La tasa del ruido ha sido del 28,5%.

3.2.2. Macroevaluación: exhaustividad y precisión

Tasa de exhaustividad: a / a + c

La exhaustividad de la búsqueda sobre Cadaqués da 5 / 5 + 3 = 0,625. La tasa de exhaustividad ha sido del 62,5%. Los valores habituales son entre 0,6 y 0,8.

Esta tasa expresa la capacidad del sistema para proporcionar lo que se quiere con un grado satisfactorio de exhaustividad. Ahora bien, con esto solo no es suficiente para evaluar la calidad, también es preciso que nos filtre lo que no necesitamos, y aquí entra la tasa de precisión.

Nota

Exhaustividad o llamada: *recall* en inglés y *rapell* en francés.

Tasa de precisión: a / a + b

La precisión de la búsqueda sobre Cadaqués da 5 / 5 + 2 = 0,714.

Resumen de las fórmulas para calcular silencio, ruido, exhaustividad y precisión.

Microevaluación		Macroevaluación	
Silencio	Ruido	Exhaustividad	Precisión
c / a + c	b / a + b	a / a + c	a / a + b

Hemos visto las tasas de silencio y ruido y las de exhaustividad y precisión, pero un análisis completo comprende el examen de los documentos, los registros de indización, las hojas de petición, las estrategias de búsqueda, las hojas de valoración de la relevancia y cualquier otra información que se pueda obtener de los usuarios que participen en el estudio. A partir de estos registros se pueden determinar las causas concretas de los errores del sistema en la recuperación.

3.3. El papel del vocabulario en la recuperación

Según Lancaster, se producen tres errores relacionados con el vocabulario:

1) falta de especificidad del lenguaje documental,

2) relaciones ambiguas, y

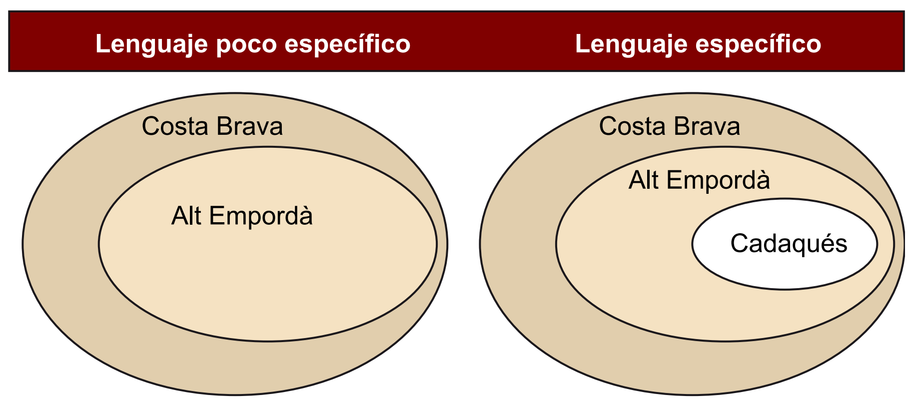
3) relaciones falsas entre términos.

3.3.1. Falta de especificidad del lenguaje documental

La falta de especificidad del lenguaje documental es la causa principal de las carencias en la recuperación y se da principalmente en el ámbito de los **lenguajes controlados.**

> **Reflexión**
>
> Recordemos que los lenguajes libres no disponen a priori de un vocabulario controlado; por lo tanto, el analista o el algoritmo del programa indizarían *Cadaqués* sin verificar si este término existe o no en una lista acotada. Los lenguajes libres son tan específicos como lo es el texto.

Figura 2. Comparación entre dos lenguajes por lo que respecta a su especificidad.



Si el lenguaje no es específico, aunque el analista quiera indexar Cadaqués, no podrá hacerlo y tendrá que recurrir a su término genérico (TG), como, por ejemplo, Alt Empordà o Costa Brava. Llegados a este punto, podemos encontrarnos con problemas tanto en la indexación como en la recuperación:

- En la **indexación**: si no hay remisiones entre términos, un analista podría indexar Costa Brava y otro analista, Alt Empordà. En cambio, si el lenguaje tiene notas de aplicación o equivalencias que remitan al término designado (tipo Cadaqués utilizar Costa Brava), todos los analistas indexarán con el mismo término, como Costa Brava, y no habrá problemas de coherencia entre ellos.
- En la **recuperación**: el usuario, que no debe conocer el lenguaje de antemano, busca Cadaqués y el sistema le devuelve 0 resultados, ya que no sabe que los documentos se han indexado con otros TG.

Si el lenguaje fuera específico y tuviera un término para Cadaqués, el lenguaje tendería a:

- Aumentar la precisión: cuando buscamos Cadaqués, recuperamos Cadaqués y no otras poblaciones de la Costa Brava o del Alt Empordà.
- Reducir la exhaustividad: solo recuperamos los documentos que tratan de Cadaqués y no los de Llançà o El Port de la Selva.

Un vocabulario específico incrementa la precisión y disminuye la exhaustividad; por el contrario, un vocabulario poco específico facilita la exhaustividad, pero reduce la precisión y aumenta la consistencia, al haber menos términos entre los que elegir.

A pesar de todo, es mejor que el lenguaje documental sea específico, es decir, **es preferible la precisión a la exhaustividad**, ya que esta se puede conseguir buscando por el TG.

En resumen:

- Un **vocabulario específico** permite una precisión alta, pero complica el hecho de conseguir una exhaustividad alta. También influye en la consistencia, ya que si los términos son muy cercanos, se puede dudar entre uno u otro.
- Un **vocabulario poco específico** facilita la búsqueda genérica y minimiza las incorrecciones de la indización y, en consecuencia, aumenta la exhaustividad, pero dificulta una precisión alta.
- Con todo, según Lancaster, es mejor un exceso de especificidad que lo contrario, ya que si queremos aumentar la exhaustividad solo hay que recurrir a los TG. En cambio, la falta de especificidad provoca que no se pueda aumentar la precisión.

3.3.2. Coordinaciones falsas

Existen dos tipos de relaciones ambiguas o falsas entre los términos: las **coordinaciones falsas** y las **relaciones incorrectas entre términos**. Ambas se producen porque las palabras en sí mismas no tienen sintaxis, especialmente si son unitérminos o términos simples.

Digamos que una coordinación es falsa cuando recuperamos documentos no pertinentes pero que contienen los términos de búsqueda que hemos pedido. La coordinación es falsa porque en el documento original los dos términos existen pero no están relacionados.

En un sistema sin sintaxis, cuanto más términos de indización haya, más alta es la probabilidad de que se recuperen coordinaciones falsas. En cambio, es menos frecuente en sistemas precoordinados, en los que hay un control más estricto. Las coordinaciones falsas se podrían solucionar si a la hora de indizar el documento se hicieran patentes las relaciones entre los términos, al menos en el caso de relacionados y no relacionados.

Ejemplo

Efectuamos una consulta sobre financiación de los archivos en Barcelona y recuperamos el documento con el que hemos iniciado este módulo, el artículo de Alice Keefer sobre repositorios digitales universitarios, por el hecho de que los términos *financiación*, *archivos* y *Barcelona* aparecen en él, aunque el artículo no hable de la financiación de los archivos barceloneses (*financiación* aparece en relación con el *open acces*, *archivos* hace referencia al autoarchivo del profesorado y *Barcelona* aparece en los datos formales del artículo).

3.3.3. Relaciones incorrectas entre términos

Las relaciones incorrectas entre términos se dan cuando el usuario busca dos términos con un tipo de relación que no es exactamente la que tiene el documento, a pesar de que constan en él.

Ejemplo extraído de Lancaster

El usuario busca por *diseño de ordenadores* y recupera documentos sobre el diseño de aviones con ordenador. Como podéis comprobar, los términos *diseño* y *ordenadores* están presentes en el documento, aunque no en el sentido de la petición.

La solución no es poner siglas de términos relacionados (TR), porque el problema es que no sabemos qué tipo de relación tienen.

La manera de solucionar este problema en un entorno poscoordinado es **asignando roles o indicadores a los descriptores**, que son códigos o cifras, verdaderos recursos (agentes) sintácticos quemarcan el rol dentro del documento.

Por ejemplo, (2) podría indicar instrumento mediano y (4) objeto, sujeto. El resultado de las indizaciones sería el siguiente:

Ejemplo de rol

Documento del diseño de aviones con ordenadores	Documento del diseño de ordenadores
Diseño Aviones (4) Ordenadores (2)	Diseño Ordenadores (4)

Para recuperar el documento inicial que quería el usuario, la búsqueda sería: Ordenadores (4) and Diseño.

Este sistema de roles es propio de las ontologías. Los lenguajes documentales actuales no llegan a especificar el rol de cada concepto, solo marcan si son términos relacionados sin especificar de qué tipo.

Otros ejemplos

Efectuamos una consulta sobre *Pintura y guerra* (en el sentido de la guerra representada en la pintura, como el *Guernica* de Picasso) y recuperamos documentos sobre pintura de guerra (maquillaje durante la guerra).

Hacemos una consulta sobre pintura catalana, en el sentido de pintores catalanes como Fortuny, Casas o Dalí, y recuperamos, además de los documentos interesantes, estos otros:

- Catalunya en la pintura (por ejemplo, la visión de Sorolla sobre el litoral catalán).
- Pintura en Catalunya (todos aquellos pintores que han pintado en Catalunya).
- Industriales de la pintura catalanes (pintores de paredes).

En resumen:

Si aumentamos la especificidad del vocabulario, nos permite representar con más matices el significado; por lo tanto, disminuye la consistencia en la indización, aumenta la precisión y baja la exhaustividad.

Resumen del aumento de la especificidad

Aumento de la especificidad	Aumenta la precisión.
	Disminuye la consistencia.
	Disminuye la exhaustividad.

Por lo que respecta a la recuperación, probablemente la estructura del lenguaje condiciona la búsqueda de manera importante. Cuanto más estructurado esté un término y cuantas más relaciones tenga, más útil resultará para construir estrategias de búsqueda (a pesar de que sean costosas).

Las coordinaciones falsas: la causa de este error es que los términos de indización se encuentran en el mismo documento pero en un contexto diferente del que busca el usuario.

Las relaciones incorrectas: la causa de este error es que el lenguaje no especifica el tipo de relación que tienen los términos entre sí.

Bibliografía

Manuales, normativas y artículos de revista

AENOR (1990). *Documentación: Directrices para el establecimiento y desarrollo de tesauros monolingües.*

AENOR (1990). *UNE-50-106 (ISO 2788-1986). Documentación: Directrices para el establecimiento y desarrollo de tesauros monolingües.*

AENOR (1994). "Norma UNE 50-113-92/1. Documentación e información. Vocabulario. Parte 1. Conceptos fundamentales". En: *Documentación: Normas fundamentales*. Madrid: AENOR.

AENOR (1996). *UNE-50-125 (ISO 5964-1985). Documentación: Directrices para la creación y desarrollo de tesauros multilingües.*

AENOR (1997). *UNE-50-125 (ISO 5964-1985). Documentación: Directrices para la creación y desarrollo de tesauros multilingües.*

AENOR (1997). *Métodos para el análisis de los documentos, determinación de su contenido y selección de los términos de indización. Norma UNE 50-121-91*. Madrid: AENOR.

AENOR (1997). "Documentación e información. Vocabulario. Parte 6: lenguajes documentales. Norma UNE-50-113/6 (ISO 5127/6)". *Revista Española de Documentación Científica* (vol. 20, núm. 4, págs. 417-436).

AENOR (2004). *Clasificación Decimal Universal (CDU): edición abreviada de la norma UNE 50001: 2000* (incluye las modificaciones de la Norma UNE 50001:2004/1M). Traducción del Master Reference File realizada por el Centro de Información y Documentación Científica (CINDOC)Adaptada por Rosa San Segundo Manuel. Madrid: AENOR.

AENOR (2004). *Clasificación Decimal Universal (CDU) de bolsillo*. Adaptada por Rosa San Segundo Manuel. Madrid: AENOR.

Aitchison, J.; Gilchrist, A.; Bawden, D. (2000). *Thesaurus construction and use: a practical manual* (4.a ed.). Chicago: Fitzroy Dearborn.

Akdag Salah, A.; Gao, C.; Suchecki, K.; Scharnhorst, A. (2010, 15 de septiembre). "The need to categorize: a comparative look at categorization in Wikipedia and the Universal Decimal Classification System" [en línea]. En: *High Throughput Humanities, a satellite meeting at the ECCS'10 European Conference on Complex Systems*. Lisboa, Portugal. <http://hth.eccs2010.eu/abstracts.htm#Akdag-Salah-te-al>

Benito, M. (1999). *El sistema de clasificación decimal universal: manual de aprendizaje*. Madrid: Taranco.

Bonilla, S. (2007). "Web Semántica y Agentes Metarrepresentacionales basados en Marcadores Discursivos" [en línea]. *Hipertext.net* (núm. 5) <http://www.hipertext.net>

Broughton, V. (2009, 29-30 de octubre). "Concepts and terms in the faceted classification: the case of UDC". En: *International UDC Seminar 2009 "Classification at a Crossroads: Multiple Directions to Usability*. La Haya.

Cañada, J. (2006). *Tipologías y estilos en el etiquetado social* [en línea]. <http://www.terremoto.net/tipologias-y-estlos-en-el-etiquetado-social/>

Codina, L.; Marcos, M. C.; Pedraza, R. (2009). *Web semántica y sistemas de información documental*. Gijón: Trea.

Currás, E. (2005). *Ontologías, taxonomía y tesauros: manual de construcción y uso*. Gijón: Trea.

Díez Carrera, C. (1999). *Técnicas y régimen de uso de la CDU (Clasificación Decimal Universal)* (134 páginas). Gijón: Trea ("Biblioteconomía y Administración Cultural", 26).

Foskett, A. (1996). *The subject approach to information*. London Library Association Publishing.

Gil Leiva, I. (2008). *Manual de indización. Teoría y práctica*. Gijón: Ediciones Trea ("Biblioteconomía y Administración Cultural", 193).

Gil Urdiciain, B. (2004). *Manual de lenguajes documentales*. Gijón: Ediciones Trea ("Biblioteconomía y Administración Cultural", 106).

Gómez Díaz, R. (2005). *La lematización en español: una aplicación para la recuperación de información*. Gijón: Trea.

Knautz, K.; Stock, W. G. (2010). "Collective indexing of emotions in videos". *Journal of Documentation* (vol. 67, núm. 6, págs. 975-994).

Lambe, P. (2007). *Organising knowledge: taxonomies, knowledge and organisational effectiveness*. Oxford: Chandos, cop.

Lancaster, F. W. (1995). *Indización y resumen: teoría y práctica*. Buenos Aires: EB Publicaciones.

Lancaster, F. W. (2002). *El control del vocabulario en la recuperación de información*. Valencia: Universitat de València.

Madalli, D. (2009, 29-30 de octubre). "Classificatory ontologies". En: *International UDC Seminar 2009 Classification at a Crossroads: Multiple Directions to Usability*. La Haya.

Maniez, J. (1992). *Los lenguajes documentales y de clasificación: concepción, construcción y utilización en los sistemas documentales*. Madrid: Pirámide / Fundación Germán Sánchez Ruipérez.

Martínez Tamayo, A. M.; Valdez, J. C. (2008). *Indización y clasificación en bibliotecas*. Buenos Aires: Alfagrama.

McIlwaine, I. C. (2003). *Clasificación Decimal Universal. Guía para uso de la CDU*. Madrid: AENOR.

Moreno, L. M.; Borgoños, M. D. (2002). *Teoría y práctica de la Clasificación decimal universal (CDU)*. Gijón: Ediciones Trea ("Biblioteconomía y Administración Cultural", 30).

Naumis, C. (2007). *Los tesauros documentales y su aplicación en la información impresa, digital y multimedia*. México: Alfagrama.

NISO Z39.19 (2003). *Guidelines for the Construction, Format, and Management of Monolingual Thesauri*.

NISO Z39.19 (2005). *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*.

Noruzzi, A. (2006). "Folksonomies: (un)controlled vocabulary?". *A Knowledge Organization* (vol. 33, núm. 4, págs. 199-203).

Olson, H. A.; Boll, J. J. (2001). *Subject Analysis in Online Catalogs*. Englewood: Libraries Unlimited.

Pinto, M. (1997). *Manual de clasificación documental*. Editorial Síntesis.

Ransom, N.; Rafferty, P. (2011). "Facets of user-assigned tags and their effectiveness in image retrieval". *Journal of Documentation* (vol. 67, núm. 6, págs. 1.038-1.066).

San Segundo, R. (2009, 29-30 de octubre). "Using MARC classification format for UDC and mappings to other KO systems for an enriched authority file". *Classification at a Crossroads: Multiple Directions to Usability*. La Haya.

Slavic, A. (2007, noviembre-diciembre). "On the nature and typology of documentary classifications and their use in a networked environment". *El Profesional de la Información* (vol. 16, núm. 6, págs. 580-589).

Slavic, A. (2008). "Use of the Universal Decimal Classification. A world-wide survey". *Journal of Documentation* (vol. 64, núm. 2).

Slavic, A.; Cordeiro, M. I.; Riesthuis, G. (2009, julio-septiembre). "El desarrollo de la Clasificación Decimal Universal: 1992-2008 y más allá" [en línea]. *Revista Española de Documentación Científica* (vol. 32, núm. 3, págs. 107-118). <http://redc.revistas.csic.es/index.php/redc/article/viewarticle/488>

Slype, van G. (1991). *Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales*. Madrid: Pirámide / Fundación Germán Sánchez Ruipérez ("Biblioteca del Libro").

Spiteri, L. (2007, septiembre). "The structure and form and folksonomy tags: the road to the public library catalogue". *Information Technology and Library*.

Trant, J. (2009). "Studying Social Tagging and Folksonomy: A Review and Framework" [en línea]. *Journal of Digital Information* (vol. 10, núm. 1). <http://dlist.sir.arizona.edu/2595/>.

UDC Consortium (2010). *Extensions and Corrections to the UDC* [en línea].<http://www.udcc.org/ec.htm>.

UDC Consortium (2010). *Master Reference File* [en línea]. <http://www.udcc.org/mrf.htm>.